Journal of Multivariate Analysis **77**, 21–53 (2001)
doi:10.1006/jmva.2000.1924, available online at http://www.idealibrary.com on IDEAL®

Exact Misclassification Probabilities for Plug-In Normal Quadratic Discriminant Functions. I. The Equal-Means Case[1]

H. Richard McFarland, III

The Monitor Group

and

Donald St. P. Richards

We consider the problem of discriminating, on the basis of random "training" samples, between two independent multivariate normal populations, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, which have a common mean vector $\boldsymbol{\mu}$ and distinct covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Using the theory of Bessel functions of the second kind of matrix argument developed by Herz (1955, *Ann. Math.* **61**, 474–523), we derive stochastic representations for the exact distributions of the "plug-in" quadratic discriminant functions for classifying a newly obtained observation. These stochastic representations involve only chi-squared and F-distributions, hence we obtain an efficient method for simulating the discriminant functions and estimating the corresponding probabilities of misclassification. For some special values of p, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ we obtain explicit formulas and inequalities for the probabilities of misclassification. We apply these results to data given by Stocks (1933, *Ann. Eugen.* **5**, 1–55) in a biometric investigation of the physical characteristics of twins, and to data provided by Rencher (1995, "Methods of Multivariate Analysis," Wiley, New York) in a study of the relationship between football helmet design and neck injuries. For each application we estimate the exact probabilities of misclassification, and in the case of Stocks' data we make extensive comparisons with previously published estimates.

AMS 1991 subject classifications: 62H10; 62H30; 62E15.

Key words and phrases: Bessel function of matrix argument; biplot; discriminant analysis; dizygotic twins; football helmet design; misclassification probability; monozygotic twins; multivariate gamma function; multivariate normality; neck injuries; stochastic representation; stochastic ordering; Wishart distribution.

[1] This research was supported in part by the National Science Foundation under Grants DMS-9401322 and DMS-9703705.

0047-259X/01 $35.00

1. INTRODUCTION

In classical multivariate discriminant analysis great attention has been paid to the problem of classifying an observation, $\mathbf{y}$, into one of two independent multivariate normal populations, $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. We refer to Anderson [1], Johnson and Wichern [8], McLachlan [13] and Muirhead [15] for extensive accounts of the theory and applications of normal discriminant analysis.

For the case in which the mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are distinct and the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are equal, the distribution theory of the resulting linear discriminant functions has been thoroughly researched, and much is known about the monotonicity properties of their misclassification probabilities (cf. Anderson [1], Muirhead [15] and McLachlan [13]).

In some applications the populations may be assumed to have a common mean vector, i.e. $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \equiv \boldsymbol{\mu}$, and distinct covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Okamoto [16] was first to consider this problem, and he gave an application to the classification of twins based on measurements of physical characteristics. Clearly, the corresponding discriminant function should reflect the assumption of equal means; however, this assumption generally causes the classification problem to become more difficult than in the case of linear discriminant analysis.

Let us denote by ϕ_1 and ϕ_2 the density functions of the populations $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, respectively. Then the *likelihood ratio classification procedure* is to classify $\mathbf{y}$ to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ if the likelihood ratio, $\phi_1(\mathbf{y})/\phi_2(\mathbf{y})$, is sufficiently large, i.e.

$$\frac{\phi_1(\mathbf{y})}{\phi_2(\mathbf{y})} > k, \tag{1.1}$$

for a suitably chosen constant k; otherwise, $\mathbf{y}$ is classified to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$.

Let q_1 and q_2 denote the known *a priori* probabilities of the first and second populations, respectively. Let $C(2 \mid 1)$ denote the cost of misclassifying $\mathbf{y}$ into $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ when, in fact, $\mathbf{y}$ belongs to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$; and let $C(1 \mid 2)$ denote the cost of misclassifying $\mathbf{y}$ into $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ when, in fact, $\mathbf{y}$ belongs to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$. It is well-known (cf. Anderson [1, p. 201]) that the choice

$$k = \frac{q_2 C(1 \mid 2)}{q_1 C(2 \mid 1)} \tag{1.2}$$

leads to the optimal, or Bayes, classification procedure under which the expected overall cost of misclassification is minimized. In particular, if $q_1 = q_2$ and $C(1 \mid 2) = C(2 \mid 1)$ then $k = 1$, and (1.1) reduces to

$$Q := \log \frac{\phi_1(\mathbf{y})}{\phi_2(\mathbf{y})} > 0. \tag{1.3}$$

It is straightforward to observe that the log-ratio in (1.3) reduces to the *equal-means discriminant function*,

$$Q_1 := \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' [\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}](\mathbf{y}-\boldsymbol{\mu}) + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}. \tag{1.4}$$

A special case of (1.4) is the *zero-means discriminant function*,

$$Q_2 := \frac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}]\,\mathbf{y} + \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}, \tag{1.5}$$

which is obtained from the case in which $\boldsymbol{\mu}$ is assumed to be known, in which case we may assume, without loss of generality, that $\boldsymbol{\mu} = \mathbf{0}$.

By omitting the term $\log|\boldsymbol{\Sigma}_2|/|\boldsymbol{\Sigma}_1|$ in (1.4) and (1.5) we obtain the *minimum distance discriminant functions*,

$$Q_3 := \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' [\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}](\mathbf{y}-\boldsymbol{\mu}) \tag{1.6}$$

and

$$Q_4 := \tfrac{1}{2}\mathbf{y}'[\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}]\,\mathbf{y}, \tag{1.7}$$

respectively.

Suppose that the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are known. If $\mathbf{y}$ belongs, say, to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ then, up to an additive constant term, the discriminant functions (1.4)–(1.7) are all of the form $\mathbf{v}'\mathbf{B}\mathbf{v}$ where $\mathbf{v}$ has a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $\mathbf{B}$ is a symmetric $p \times p$ matrix. Then we can show, by the standard method of reducing $\mathbf{B}$ to diagonal form through an orthogonal transformation, that each Q_j is equal in distribution to a linear combination of independent chi-squared random variables.

In many practical applications the mean vector $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are unknown, and it is necessary to estimate them. Suppose we collect two mutually independent random "training" samples, $\mathbf{y}_1^{(1)}, ..., \mathbf{y}_{N_1}^{(1)}$ and $\mathbf{y}_1^{(2)}, ..., \mathbf{y}_{N_2}^{(2)}$, drawn from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, respectively. Let $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ denote the corresponding sample means and $\bar{\mathbf{y}} = (N_1\bar{\mathbf{y}}_1 + N_2\bar{\mathbf{y}}_2)/(N_1 + N_2)$ be the pooled estimate of $\boldsymbol{\mu}$; hence $\bar{\mathbf{y}}$ is an unbiased estimator of $\boldsymbol{\mu}$. For $g = 1, 2$, let

$$\mathbf{S}_g = \frac{1}{N_g - 1}\sum_{i=1}^{N_g}(\mathbf{y}_i^{(g)} - \bar{\mathbf{y}}_g)(\mathbf{y}_i^{(g)} - \bar{\mathbf{y}}_g)'$$

be the sample covariance matrix corresponding to the gth training sample; hence $\mathbf{S}_g$ is an unbiased estimator of $\boldsymbol{\Sigma}_g$, $g=1, 2$. Finally, let $\mathbf{y}$ be a new observation which is known to belong either to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ or to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$, and which is independent of both training samples. Then sample analogs of the discriminant functions (1.4)–(1.7) are

$$\hat{Q}_1 = c_1(\mathbf{y}-\bar{\mathbf{y}})' \, [\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}](\mathbf{y}-\bar{\mathbf{y}}) + c_2 \log \frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}, \tag{1.8}$$

$$\hat{Q}_2 = c_1 \mathbf{y}'[\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}] \, \mathbf{y} + c_2 \log \frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}, \tag{1.9}$$

$$\hat{Q}_3 = c_1(\mathbf{y}-\bar{\mathbf{y}})' \, [\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}](\mathbf{y}-\bar{\mathbf{y}}), \tag{1.10}$$

$$\hat{Q}_4 = c_1 \mathbf{y}'[\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}] \, \mathbf{y}, \tag{1.11}$$

respectively, where $c_1 = c_2 = \frac{1}{2}$. These criteria are usually called the "plug-in" discriminant functions because they retain a form similar to (1.4)–(1.7), but with unbiased estimators "plugged-in" for the corresponding parameters.

The distribution theory of these statistics is far more complicated than their counterparts (1.4)–(1.7). Okamoto [16] first studied this problem, and derived the asymptotic distributions of approximations to (1.8) and (1.10). More precisely, Okamoto treated the term $\log |\mathbf{S}_2|/|\mathbf{S}_1|$ as a constant and then derived an asymptotic expansion for the distribution of a statistic which *approximates* (1.10); even with these simplifications, the resulting expressions are recondite (cf. Okamoto [16], Siotani [18]).

Since the appearance of Okamoto's article [16], we have found in the literature distributional results pertaining only to *asymptotic expansions* for the distributions of (1.8)–(1.11), i.e. under the assumption that $N_1, N_2 \to \infty$. Moreover, in all instances, these results were derived either under special assumptions about the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, or as the result of an asymptotic expansion. The most recent results, due to Marco *et al.* [10], developed expressions for the asymptotic misclassification probabilities under a uniform covariance assumption due to Bartlett and Please [2], *viz.*,

$$\boldsymbol{\Sigma}_g = \sigma_g^2\{(1-\rho_g)\,\mathbf{I} + \rho_g \mathbf{1}_p \mathbf{1}_p'\}, \qquad g=1, 2, \tag{1.12}$$

where $\mathbf{1}_p = (1, ..., 1)'$ is a $p \times 1$ vector. Marco *et al.* [10] assumed the parameters in (1.12) were unknown, provided estimates for those parameters, derived asymptotic expansions for the distributions of the corresponding discriminant functions, and gave an application to the well-known data set of Stocks [19] on the problem of classifying twin pairs of children into monozygotic or dizygotic populations.

Despite the great attention paid to the asymptotic distributions of the discriminant functions (1.8)–(1.11) it is often the case that the sample sizes are only moderately large. For example, Rencher [17, pp. 306–307] provided data obtained in a study of potential links between football helmet design and players' neck injuries; in this example, which we discuss in greater detail in Section 5, $(N_1+N_2)/p=10$. Also included in Section 5 is an analysis of the well-known data set provided by Stocks [19]; in this example, we have $(N_1+N_2)/p=9$. In these applications, the asymptotic theory obviously cannot be expected to provide accurate approximations to the probabilities of misclassification. These remarks indicate clearly that there remains the need for an investigation of the *exact* distribution of the plug-in discriminant functions. We note also that, although the results of Bowker [3] on exact stochastic representations for Anderson's linear discriminant function have long been available, no such results have been obtained before now for any quadratic discriminant function.

In this paper, which is based primarily on the dissertation of McFarland [12], we derive representations for the exact distributions, and corresponding misclassification probabilities, of the plug-in quadratic discriminant functions (1.8)–(1.11). For each of the statistics (1.8)–(1.11), we obtain a stochastic representation in terms of classical *univariate* random variables. In our development we make no assumptions about the constants c_1 and c_2 in (1.8)–(1.11).

Our approach to deriving these stochastic representations is by way of a direct analysis of the corresponding characteristic functions. By application of results of Herz [6], we obtain the characteristic functions of the discriminant functions (1.8)–(1.11) in terms of the Bessel functions of matrix argument of the second kind. Then we apply some detailed properties of these Bessel functions to simplify the problem. By several transformations of variables, we deduce stochastic representations for the exact distributions of the discriminant functions (1.8)–(1.11).

In the case of $\hat{Q}_1$ and $\hat{Q}_2$, our stochastic representations involve $2p+1$ independent random variables each having a chi-squared or an F-distribution. In the case of $\hat{Q}_3$ and $\hat{Q}_4$ the corresponding number of independent random variables is $p+1$, and all are chi-squared or F-distributed. Therefore for all four plug-in discriminant functions, Monte Carlo simulation of the exact distribution functions starting from the new stochastic representations will be more efficient than a direct Monte Carlo simulation of the discriminant statistic itself.

Let us now outline the results in the paper. In Section 2 we provide some preliminaries relating to the multivariate gamma function and the Bessel functions of matrix argument of the second kind. In Section 3 we derive stochastic representations for the exact distributions of (1.8)–(1.11). In Section 4 we apply the stochastic representations derived in Section 3 to study the corresponding probabilities of misclassification. We obtain general formulas for these misclassification probabilities and, in some special cases, we apply stochastic bounds for the stochastic representations to derive inequalities for the misclassification probabilities.

In Section 5, we apply the results of Sections 3 and 4 to estimate the misclassification probabilities associated with Stocks' [19] measurements on the physical characteristics of twins and Rencher's [17] data on football players' head sizes. These data sets were chosen from the discriminant analysis literature to illustrate the behavior of the distributions of the discriminant functions $\hat{Q}_j$, $j=1, ..., 4$. For each of these data sets, we apply Mardia's statistic [11] to test for multivariate normality. We employ the method of biplots (cf. Gabriel [5], Khattree and Naik [9]) as a graphical exploratory tool to display multivariate relationships between the observations, the variables, and the population groups *via* two-dimensional plots. From these biplots, we construct ellipses to illustrate graphically the population distributions for the groups. These ellipses depict the relative covariance structures (i.e. shape, volume, and orientation) for each of the groups. In this way, the biplots provide us with a graphical tool for interpreting the results obtained from Monte Carlo simulation of the distributions of the corresponding discriminant functions.

2. PRELIMINARIES

Throughout we denote by $\mathbf{\Pi}_g$ the population $N_p(\boldsymbol{\mu}, \mathbf{\Sigma}_g)$, $g=1, 2$. We also denote by $\lambda_1, ..., \lambda_p$ the eigenvalues of the matrix $\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$.

Let $\mathbf{y}_1^{(1)}, ..., \mathbf{y}_{N_1}^{(1)}$ and $\mathbf{y}_1^{(2)}, ..., \mathbf{y}_{N_2}^{(2)}$ be independent random training samples from $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, respectively with $N_1, N_2 > p$. If $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ denote the corresponding sample means, then the pooled estimate of $\boldsymbol{\mu}$ is $\bar{\mathbf{y}} = (N_1\bar{\mathbf{y}}_1 + N_2\bar{\mathbf{y}}_2)/(N_1+N_2)$, and for $g=1, 2$, we let

$$\mathbf{A}_g = \sum_{i=1}^{N_g} (\mathbf{y}_i^{(g)} - \bar{\mathbf{y}}_g)(\mathbf{y}_i^{(g)} - \bar{\mathbf{y}}_g)', \tag{2.1}$$

the matrix of sums of squares and products. If we let $n_g = N_g - 1$, $g=1, 2$, then an unbiased estimate of the covariance matrix $\mathbf{\Sigma}_g$ is $\mathbf{S}_g = n_g^{-1}\mathbf{A}_g$. It is well-known that $\mathbf{A}_g$ has a Wishart distribution, denoted $\mathbf{A}_g \stackrel{d}{=} W_p(n_g, \mathbf{\Sigma}_g)$.

DEFINITION 2.1. Let $\delta \in \mathbb{C}$ such that $\mathrm{Re}(\delta) > \frac{1}{2}(p-1)$. Then the *multivariate gamma function* is defined as

$$\Gamma_p(\delta) := \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(\delta - \tfrac{1}{2}(j-1)) \tag{2.2}$$

(cf. Muirhead [15, pp. 61–63]).

The following result is a straightforward consequence of Definition 2.1.

LEMMA 2.2. *For* $\delta \in \mathbb{C}$ *such that* $\mathrm{Re}(\delta) > \frac{1}{2}(p-1)$,

$$\frac{\Gamma_p(\delta)}{\Gamma(\delta - \frac{1}{2}(p-1))} = \pi^{1/2(p-1)} \Gamma_{p-1}(\delta). \tag{2.3}$$

LEMMA 2.3. *Let* $\mathrm{F}_1, ..., \mathrm{F}_{p-1}$ *be independent F-distributed random variables, where* F_j *has degrees of freedom* $(N_2 - j, N_1 - j)$, $j = 1, ..., p-1$. *Then, for* $t \in \mathbb{R}$,

$$\mathsf{E}[\mathrm{e}^{it \sum_{j=1}^{p-1} \log((N_2 - j)/(N_1 - j)\,\mathrm{F}_j)}] = \frac{\Gamma_{p-1}(\frac{1}{2}n_1 - it)\, \Gamma_{p-1}(\frac{1}{2}n_2 + it)}{\Gamma_{p-1}(\frac{1}{2}n_1)\, \Gamma_{p-1}(\frac{1}{2}n_2)}. \tag{2.4}$$

Proof. Let $\mathrm{U} \stackrel{d}{=} \chi_n^2$, a chi-squared distribution on n degrees of freedom. It is well-known that

$$\mathsf{E}[\mathrm{e}^{it \log(\mathrm{U}/2)}] = 2^{-it} \mathsf{E}[\mathrm{U}^{it}] = \frac{\Gamma(\frac{1}{2}n + it)}{\Gamma(\frac{1}{2}n)}.$$

It follows from (2.2) that, for $n \geqslant p$,

$$\begin{aligned} \frac{\Gamma_{p-1}(\frac{1}{2}n + it)}{\Gamma_{p-1}(\frac{1}{2}n)} &= \prod_{j=1}^{p-1} \frac{\Gamma(\frac{1}{2}(n-j+1) + it)}{\Gamma(\frac{1}{2}(n-j+1))} \\ &= \prod_{j=1}^{p-1} \mathsf{E}[\mathrm{e}^{it \log(\mathrm{U}_j/2)}] \\ &= \mathsf{E}[\mathrm{e}^{it \sum_{j=1}^{p-1} \log(\mathrm{U}_j/2)}], \end{aligned}$$

where $\mathrm{U}_1, ..., \mathrm{U}_{p-1}$ are mutually independent random variables, with $\mathrm{U}_j \stackrel{d}{=} \chi^2_{n+1-j}$, $j = 1, ..., p-1$. Therefore

$$\begin{aligned} \frac{\Gamma_{p-1}(\frac{1}{2}n_1 - it)\, \Gamma_{p-1}(\frac{1}{2}n_2 + it)}{\Gamma_{p-1}(\frac{1}{2}n_1)\, \Gamma_{p-1}(\frac{1}{2}n_2)} &= \mathsf{E}[\mathrm{e}^{-it \sum_{j=1}^{p-1} \log(\mathrm{U}_{1j}/2)}\, \mathrm{e}^{it \sum_{j=1}^{p-1} \log(\mathrm{U}_{2j}/2)}] \\ &= \mathsf{E}[\mathrm{e}^{it \sum_{j=1}^{p-1} \log \mathrm{U}_{2j}/\mathrm{U}_{1j}}], \end{aligned} \tag{2.5}$$

where the random variables U_{gj}, $g = 1, 2$, $j = 1, ..., p-1$, are mutually independent, and $\mathrm{U}_{gj} \stackrel{d}{=} \chi^2_{N_g - j}$. Then

$$\frac{\mathrm{U}_{2j}}{\mathrm{U}_{1j}} \stackrel{d}{=} \frac{(N_2 - j)}{(N_1 - j)}\, \mathrm{F}_j, \tag{2.6}$$

where $\mathrm{F}_1, ..., \mathrm{F}_{p-1}$ are mutually independent and F-distributed with degrees of freedom $(N_2-1, N_1-1), ..., (N_2-p+1, N_2-p+1)$, respectively. Substituting (2.6) into (2.5), then, completes the proof. ∎

It is well-known (cf. Anderson [1, p. 264]) that if U_1 and U_2 are independent random variables, $\mathrm{U}_1 \stackrel{d}{=} \chi^2_{N-1}$ and $\mathrm{U}_2 \stackrel{d}{=} \chi^2_{N-2}$, then $4\mathrm{U}_1\mathrm{U}_2 \stackrel{d}{=} \chi^2_{2N-4}$. By applying this result to the random variables U_{gj}, $j=1, ..., p-1$, $g=1, 2$, in the proof of Lemma 2.3, we obtain a result in which the number of F-distributed random variables in (2.4) is reduced by one-half.

COROLLARY 2.4. *Suppose p is odd, $p=2r+1$. Then for $t \in \mathbb{R}$,*

$$\mathsf{E}\left[e^{it \sum_{j=1}^{r} \log((N_2-j)/(N_1-j)\,\mathrm{F}_j)}\right] = \frac{\Gamma_{2r}(\frac{1}{2}n_1 - it)\,\Gamma_{2r}(\frac{1}{2}n_2 + it)}{\Gamma_{2r}(\frac{1}{2}n_1)\,\Gamma_{2r}(\frac{1}{2}n_2)}, \tag{2.7}$$

where the random variables $\mathrm{F}_1, ..., \mathrm{F}_r$ are independent and F-distributed, and F_j has degrees of freedom $(2N_2-4j, 2N_1-4j)$, $j=1, ..., r$.

If p is even, $p=2r$, then

$$\mathsf{E}\left[e^{it\{\log((N_2-2r+1)/(N_1-2r+1)\,\mathrm{F}_r)) + \sum_{j=1}^{r-1} \log((N_2-j)/(N_1-j)\,\mathrm{F}_j))\}}\right]$$
$$= \frac{\Gamma_{2r-1}(\frac{1}{2}n_1 - it)\,\Gamma_{2r-1}(\frac{1}{2}n_2 + it)}{\Gamma_{2r-1}(\frac{1}{2}n_1)\,\Gamma_{2r-1}(\frac{1}{2}n_2)}, \tag{2.8}$$

where the random variables $\mathrm{F}_1, ..., \mathrm{F}_r$ are independent and F-distributed; F_j has degrees of freedom $(2N_2-4j, 2N_1-4j)$, $j=1, ..., r-1$, and F_r has degrees of freedom $(N_2-2r+1, 2N_1-2r+1)$.

We now introduce the Bessel functions of matrix argument of the second kind. We denote by $\{\mathbf{\Lambda} > \mathbf{0}\}$ the space of $p \times p$, positive-definite, symmetric matrices $\mathbf{\Lambda}$, and we denote by $d\mathbf{\Lambda}$ the Lebesgue measure on the space $\{\mathbf{\Lambda} > \mathbf{0}\}$.

DEFINITION 2.5 (Herz [6]). Let $\mathbf{Z}$ be a $p \times p$ complex symmetric matrix and $\delta \in \mathbb{C}$. Then the *Bessel function of matrix argument of the second kind* is

$$\mathrm{B}_\delta^{(p)}(\mathbf{Z}) = \int_{\mathbf{\Lambda} > \mathbf{0}} e^{-\operatorname{tr}(\mathbf{\Lambda Z} + \mathbf{\Lambda}^{-1})}\, |\mathbf{\Lambda}|^{\delta - 1/2(p+1)}\, d\mathbf{\Lambda}. \tag{2.9}$$

If $\mathrm{Re}(\mathbf{Z}) = \mathbf{0}$, i.e. $\mathbf{Z}$ is purely imaginary, then (2.9) is absolutely convergent if and only if $\mathrm{Re}(\delta) < -\frac{1}{2}(p-1)$ (Herz [6, p. 506]). If $\mathrm{Re}(\mathbf{Z}) > \mathbf{0}$ then the integral (2.9) converges absolutely for all $\delta \in \mathbb{C}$, and then

$$\mathrm{B}_\delta^{(p)}(\mathbf{Z}) = \int_{\mathbf{\Lambda} > \mathbf{0}} e^{\operatorname{tr}(-\mathbf{\Lambda} - \mathbf{\Lambda}^{-1}\mathbf{Z})}\, |\mathbf{\Lambda}|^{-\delta-(p+1)/2}\, d\mathbf{\Lambda}. \tag{2.10}$$

It also follows from (2.9) and (2.10) that, for any $p \times p$ orthogonal matrix $\mathbf{H}$,

$$\mathrm{B}_\delta^{(p)}(\mathbf{H}\mathbf{Z}\mathbf{H}^{-1}) = \mathrm{B}_\delta^{(p)}(\mathbf{Z});$$

therefore, if $\mathbf{Z}$ is a real or imaginary symmetric matrix then $\mathrm{B}_\delta^{(p)}(\mathbf{Z})$ depends only on the eigenvalues of $\mathbf{Z}$.

Substituting $p=1$ in Definition 2.5, we obtain

$$\mathrm{B}_\delta^{(1)}(z) = \int_0^\infty \mathrm{e}^{-zx - x^{-1}} x^{\delta - 1}\, dx, \qquad z \in \mathbb{C}, \tag{2.11}$$

which is absolutely convergent for all $z \in \mathbb{C}$ such that $\mathrm{Re}(z) = 0$ whenever $\mathrm{Re}(\delta) < 0$. In the classical case in which $p = 1$ we have

$$\mathrm{B}_\delta^{(1)}(z^2) = \frac{\pi}{\sin(\pi\delta)}\, z^{-\delta}[\mathrm{I}_{-\delta}(2z) - \mathrm{I}_\delta(2z)],$$

where $\mathrm{I}_\delta(\cdot)$ is the modified Bessel function of the first kind of order δ (cf. Watson [20, p. 78, 3.7(6)]).

The following result, due to Herz [6, Theorem 5.10, p. 509], allows us in some instances to reduce the dimensionality of the Bessel functions of matrix argument.

Lemma 2.6 (Herz [6]). *Suppose $\mathbf{Z}$ is a $p \times p$ real or imaginary symmetric matrix of rank k, $k < p$, and let $\tilde{\mathbf{Z}}$ be any $k \times k$ symmetric matrix whose eigenvalues are the k non-zero eigenvalues of $\mathbf{Z}$. If* $\mathrm{Re}(\delta) < -\frac{1}{2}(p-k-1)$ *then*

$$\mathrm{B}_{-\delta}^{(p)}(\mathbf{Z}) = \frac{\Gamma_p(\delta)}{\Gamma_k(\delta - \frac{1}{2}(p-k))}\, \mathrm{B}_{-\delta + (p-k)/2}^{(k)}(\tilde{\mathbf{Z}}).$$

Let $n \geqslant p$, $\boldsymbol{\Sigma} > 0$ and suppose that the symmetric $p \times p$ random matrix $\mathbf{A}$ has a Wishart distribution, $\mathbf{A} \stackrel{d}{=} \mathrm{W}_p(n, \boldsymbol{\Sigma})$, with density function

$$\frac{1}{|2\boldsymbol{\Sigma}|^{1/2n}\, \Gamma_p(\frac{1}{2}n)}\, |\mathbf{A}|^{1/2(n-p-1)}\, \mathrm{e}^{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{A})/2}, \qquad \mathbf{A} > \mathbf{0}. \tag{2.12}$$

The following result provides a representation for the joint characteristic function of $\mathbf{y}'\mathbf{A}^{-1}\mathbf{y}$ and $\log|\mathbf{A}|$, where $\mathbf{y} \in \mathbb{R}^p$, in terms of the Bessel functions $B_\delta^{(1)}$.

LEMMA 2.7. *Let* $\mathbf{A} \overset{d}{=} W_p(n, \boldsymbol{\Sigma})$, *where* $n = N-1$ *with* $N > p$; $\mathbf{y} \in \mathbb{R}^p$ *be a fixed vector*; *and* $t_1, t_2 \in \mathbb{R}$. *Then*

$$\mathsf{E}[e^{i(t_1 \mathbf{y}'\mathbf{A}^{-1}\mathbf{y} + t_2 \log |\mathbf{A}|)}] = \frac{|2\boldsymbol{\Sigma}|^{it_2} \pi^{(p-1)/2} \Gamma_{p-1}(\frac{1}{2}n + it_2)}{\Gamma_p(\frac{1}{2}n)} B^{(1)}_{-it_2-(N-p)/2}\left(-\frac{1}{2} it_1 \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right). \tag{2.13}$$

In particular, for $t \in \mathbb{R}$,

$$\mathsf{E}[e^{it\mathbf{y}'\mathbf{A}^{-1}\mathbf{y}}] = \frac{1}{\Gamma(\frac{1}{2}(N-p))} B^{(1)}_{-(N-p)/2}\left(-\frac{1}{2} it\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right). \tag{2.14}$$

Proof. By (2.12), we have

$$\begin{aligned}\mathsf{E}[e^{i(t_1 \mathbf{y}'\mathbf{A}^{-1}\mathbf{y} + t_2 \log |\mathbf{A}|)}] &= \mathsf{E}[|\mathbf{A}|^{it_2} e^{it_1\mathbf{y}'\mathbf{A}^{-1}\mathbf{y}}] \\ &= \int_{\mathbf{A}>\mathbf{0}} |\mathbf{A}|^{it_2} e^{it_1\mathbf{y}'\mathbf{A}^{-1}\mathbf{y}} \frac{|\mathbf{A}|^{(n-p-1)/2} e^{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{A})/2}}{|2\boldsymbol{\Sigma}|^{n/2} \Gamma_p(\frac{1}{2}n)} d\mathbf{A}.\end{aligned} \tag{2.15}$$

Next we make the transformation $\boldsymbol{\Lambda} = (2\boldsymbol{\Sigma})^{1/2} \mathbf{A}^{-1} (2\boldsymbol{\Sigma})^{1/2}$; it is well-known that the Jacobian of this transformation is $|2\boldsymbol{\Sigma}|^{(p+1)/2} |\boldsymbol{\Lambda}|^{-(p+1)}$ (cf. Anderson [1, pp. 255, 268]). Then the right-hand side of (2.15) becomes

$$\frac{|2\boldsymbol{\Sigma}|^{it_2}}{\Gamma_p(\frac{1}{2}n)} \int_{\boldsymbol{\Lambda}>\mathbf{0}} |\boldsymbol{\Lambda}|^{-it_2-(n+p+1)/2} e^{\mathrm{tr}(it_1\boldsymbol{\Sigma}^{-1/2}\mathbf{y}\mathbf{y}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Lambda})/2} e^{-\mathrm{tr}(\boldsymbol{\Lambda}^{-1})} d\boldsymbol{\Lambda}. \tag{2.16}$$

It now follows from Definition 2.5 and (2.16) that

$$\mathsf{E}[e^{i(t_1 \mathbf{y}'\mathbf{A}^{-1}\mathbf{y} + t_2 \log |\mathbf{A}|)}] = \frac{|2\boldsymbol{\Sigma}|^{it_2}}{\Gamma_p(\frac{1}{2}n)} B^{(p)}_{-it_2-n/2}\left(-\frac{1}{2} it_1 \boldsymbol{\Sigma}^{-1/2}\mathbf{y}\mathbf{y}'\boldsymbol{\Sigma}^{-1/2}\right). \tag{2.17}$$

Since the matrix $\boldsymbol{\Sigma}^{-1/2}\mathbf{y}\mathbf{y}'\boldsymbol{\Sigma}^{-1/2}$ is of rank one and $\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ is its single non-zero eigenvalue then, by an application of Lemma 2.6, we may reduce the Bessel function $B^{(p)}$ in (2.17) to a one-dimensional Bessel function $B^{(1)}$; then we obtain

$$\mathsf{E}[e^{i(t_1 \mathbf{y}'\mathbf{A}^{-1}\mathbf{y} + t_2 \log |\mathbf{A}|)}] = |2\boldsymbol{\Sigma}|^{it_2} \frac{\Gamma_p(\frac{1}{2}n + it_2)}{\Gamma_p(\frac{1}{2}n)\, \Gamma(it_2 + \frac{1}{2}(n-p+1))} B^{(1)}_{-it_2-(N-p)/2}\left(-\frac{1}{2} it_1 \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\right). \tag{2.18}$$

Applying Lemma 2.2 to combine the multivariate gamma factors in (2.18), we obtain (2.13).

Finally, (2.14) is obtained by setting $t_1 = t$ and $t_2 = 0$ in (2.13) and applying Lemma 2.2. ∎

LEMMA 2.8. *Let* $V_1, ..., V_p$ *be independent, identically distributed (i.i.d.)* χ_1^2 *random variables;* $\lambda_1, ..., \lambda_p$ *denote the eigenvalues of* $\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1$; $\delta_g \in \mathbb{C}$ *with* $\mathrm{Re}(\delta_g) < 0$, $g = 1, 2$; *and* $t_1, t_2 \in \mathbb{R}$. *If* $\mathbf{y} \overset{d}{=} N_p(\mathbf{0}, \mathbf{\Sigma}_1)$ *then*

$$\mathsf{E}_{\mathbf{y}}[B_{\delta_1}^{(1)}(it_1 \mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y})\, \mathrm{B}_{\delta_2}^{(1)}(it_2 \mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y})]$$
$$= \int_0^\infty \int_0^\infty \mathsf{E}[\mathrm{e}^{-i\sum_{j=1}^p (t_1 x_1 + t_2 x_2 \lambda_j) \mathrm{V}_j}] \prod_{g=1}^2 \mathrm{e}^{-x_g^{-1}} x_g^{\delta_g - 1}\, dx_g. \tag{2.19}$$

Analogously, if $\mathbf{y} \overset{d}{=} \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}_2)$ *then*

$$\mathsf{E}_{\mathbf{y}}[B_{\delta_1}^{(1)}(it_1 \mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y})\, \mathrm{B}_{\delta_2}^{(1)}(it_2 \mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y})]$$
$$= \int_0^\infty \int_0^\infty \mathsf{E}[\mathrm{e}^{-i\sum_{j=1}^p (t_1 x_1 \lambda_j^{-1} + t_2 x_2) \mathrm{V}_j}] \prod_{g=1}^2 \mathrm{e}^{-x_g^{-1}} x_g^{\delta_g - 1}\, dx_g. \tag{2.20}$$

Proof. We begin by changing the one-dimensional Bessel functions into their integral forms by applying (2.11) of Definition 2.5, *viz.*

$$B_{\delta_g}^{(1)}(it_g \mathbf{y}'\mathbf{\Sigma}_g^{-1}\mathbf{y}) = \int_0^\infty \mathrm{e}^{-it_g \mathbf{y}'\mathbf{\Sigma}_g^{-1}\mathbf{y} x_g - x_g^{-1}} x_g^{\delta_g - 1}\, dx_g, \qquad g = 1, 2.$$

Since $\mathrm{Re}(\delta_g) < 0$, $g = 1, 2$, then these integrals converge absolutely. By Fubini's theorem we may interchange the expectation and integrals, obtaining

$$\mathsf{E}_{\mathbf{y}}[B_{\delta_1}^{(1)}(it_1 \mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y})\, B_{\delta_2}^{(1)}(it_2 \mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y})]$$
$$= \mathsf{E}_{\mathbf{y}} \int_0^\infty \mathrm{e}^{-it_1 \mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y} x_1 - x_1^{-1}} x_1^{\delta_1 - 1}\, dx_1 \int_0^\infty \mathrm{e}^{-it_2 \mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y} x_2 - x_2^{-1}} x_2^{\delta_2 - 1}\, dx_2$$
$$= \int_0^\infty \int_0^\infty \mathsf{E}_{\mathbf{y}}[\mathrm{e}^{-\mathbf{y}'(it_1 x_1 \mathbf{\Sigma}_1^{-1} + it_2 x_2 \mathbf{\Sigma}_2^{-1})\mathbf{y}}] \prod_{g=1}^2 \mathrm{e}^{-x_g^{-1}} x_g^{\delta_g - 1}\, dx_g. \tag{2.21}$$

Suppose $\mathbf{y} \overset{d}{=} \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}_1)$; then it is well-known that

$$\mathsf{E}_{\mathbf{y}}[\mathrm{e}^{-i\mathbf{y}'\mathbf{B}\mathbf{y}}] = |\mathbf{I} + 2i\mathbf{B}\mathbf{\Sigma}_1|^{-1/2} \tag{2.22}$$

for any $p \times p$ symmetric matrix $\mathbf{B}$. Therefore

$$\begin{aligned}
\mathsf{E}_{\mathbf{y}}[\mathrm{e}^{-i\mathbf{y}'(it_1x_1\boldsymbol{\Sigma}_1^{-1}+t_2x_2\boldsymbol{\Sigma}_2^{-1})\mathbf{y}}] &= |\mathbf{I}+2i(t_1x_1\boldsymbol{\Sigma}_1^{-1}+t_2x_2\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\Sigma}_1|^{-1/2}, \\
&= \prod_{j=1}^{p} [1+2i(t_1x_1+t_2x_2\lambda_j)]^{-1/2} \\
&= \prod_{j=1}^{p} \mathsf{E}[\mathrm{e}^{-i(t_1x_1+t_2x_2\lambda_j)\,\mathrm{V}_j}] \\
&= \mathsf{E}[\mathrm{e}^{-i\sum_{j=1}^{p}(t_1x_1+t_2x_2\lambda_j)\,\mathrm{V}_j}], \qquad (2.23)
\end{aligned}$$

where the random variables $\mathrm{V}_1, ..., \mathrm{V}_p$ are i.i.d. χ_1^2. Substituting (2.23) into (2.21), we obtain (2.19).

Similarly, if $\mathbf{y} \overset{d}{=} \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_2)$ then the analog of (2.23) is

$$\begin{aligned}
\mathsf{E}_{\mathbf{y}}[\mathrm{e}^{-i\mathbf{y}'(t_1x_1\boldsymbol{\Sigma}_1^{-1}+t_2x_2\boldsymbol{\Sigma}_2^{-1})\,\mathbf{y}}] &= |\mathbf{I}+2i[t_1x_1\boldsymbol{\Sigma}_1^{-1}+t_2x_2\boldsymbol{\Sigma}_2^{-1}]\boldsymbol{\Sigma}_2|^{-1/2} \\
&= \mathsf{E}[\mathrm{e}^{-i\sum_{j=1}^{p}(t_1x_1\lambda_j^{-1}+t_2x_2)\,\mathrm{V}_j}]. \qquad (2.24)
\end{aligned}$$

Substituting (2.24) into (2.21), we obtain (2.20). ▮

By an argument similar to the proof of the preceding result, we obtain the following.

Lemma 2.9. *Let* $\mathrm{V}_1, ..., \mathrm{V}_p$ *be i.i.d.* χ_1^2 *random variables;* $\delta_g \in \mathbb{C}$ *with* $\mathrm{Re}(\delta_g) < 0$, $g = 1, 2$; $t_1, t_2 \in \mathbb{R}$; *and* $\tau_1, \tau_2 > 0$. *If* $\mathbf{y} \overset{d}{=} \mathrm{N}_p(\mathbf{0}, \tau_1\boldsymbol{\Sigma}_1 + \tau_2\boldsymbol{\Sigma}_2)$, *then*

$$\begin{aligned}
&\mathsf{E}_{\mathbf{y}}[B_{\delta_1}^{(1)}(it_1\mathbf{y}'\boldsymbol{\Sigma}_1^{-1}\mathbf{y})\,\mathrm{B}_{\delta_2}^{(1)}(it_2\mathbf{y}'\boldsymbol{\Sigma}_2^{-1}\mathbf{y})] \\
&\quad = \int_0^\infty \int_0^\infty \mathsf{E}[\mathrm{e}^{-i\sum_{j=1}^{p}[t_1x_1(\tau_1+\tau_2\lambda_j^{-1})+t_2x_2(\tau_1\lambda_j+\tau_2)]\,\mathrm{V}_j}] \prod_{g=1}^{2} \mathrm{e}^{-x_g^{-1}} x_g^{\delta_g-1}\, dx_g. \qquad (2.25)
\end{aligned}$$

3. STOCHASTIC REPRESENTATIONS

First, we derive stochastic representations for the distributions of the zero-means discriminant functions $\hat{Q}_2 = c_1\mathbf{y}'(\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1})\mathbf{y} + c_2 \log |\mathbf{S}_1^{-1}\mathbf{S}_2|$ and $\hat{Q}_4 = c_1\mathbf{y}'(\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1})\mathbf{y}$, where $\mathbf{y}$, $\mathbf{S}_1$, and $\mathbf{S}_2$ are mutually independent; $\mathbf{A}_g = n_g\mathbf{S}_g \overset{d}{=} \mathrm{W}_p(n_g, \boldsymbol{\Sigma}_g)$, $g = 1, 2$; and $\mathbf{y}$ is an observation either from $\boldsymbol{\Pi}_1$ or $\boldsymbol{\Pi}_2$, where $\boldsymbol{\Pi}_g$ is the population $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_g)$, $g = 1, 2$.

Theorem 3.1. *Let* $\mathrm{T}_1, \mathrm{T}_2, \mathrm{V}_1, ..., \mathrm{V}_p$, *and* $\mathrm{F}_1, ..., \mathrm{F}_{p-1}$ *be mutually independent random variables where* $\mathrm{T}_g \overset{d}{=} \chi^2_{N_g-p}$, $g = 1, 2$; $\mathrm{V}_j \overset{d}{=} \chi_1^2$, $j = 1, ..., p$;

and F_j *be F-distributed with* (N_2-j, N_1-j) *degrees of freedom,* $j=1, ..., p-1$. *If* $\mathbf{y} \in \mathbf{\Pi}_1$ *then*

$$\hat{Q}_2 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left(\frac{n_2}{\mathrm{T}_2}\lambda_j - \frac{n_1}{\mathrm{T}_1}\right) \mathrm{V}_j + c_2 \left[\log\left(\frac{n_1^p \mathrm{T}_2}{n_2^p \mathrm{T}_1}\right) - \log |\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1| + \sum_{j=1}^{p-1} \log\left(\frac{N_2-j}{N_1-j}\,\mathrm{F}_j\right)\right], \tag{3.1}$$

and if $\mathbf{y} \in \mathbf{\Pi}_2$ *then*

$$\hat{Q}_2 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left(\frac{n_2}{\mathrm{T}_2} - \frac{n_1}{\mathrm{T}_1}\frac{1}{\lambda_j}\right) \mathrm{V}_j + c_2 \left[\log\left(\frac{n_1^p \mathrm{T}_2}{n_2^p \mathrm{T}_1}\right) - \log |\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1| + \sum_{j=1}^{p-1} \log\left(\frac{N_2-j}{N_1-j}\,\mathrm{F}_j\right)\right]. \tag{3.2}$$

Proof. Since $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{y}$ are mutually independent then the characteristic function of $\hat{Q}_2$ is

$$\begin{aligned}
\mathsf{E}[\mathrm{e}^{it\hat{Q}_2}] &= \mathsf{E}[\mathrm{e}^{itc_1\mathbf{y}'[\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}]\mathbf{y}}\,\mathrm{e}^{itc_2\log|\mathbf{S}_2|/|\mathbf{S}_1|}] \\
&= \mathsf{E}_{\mathbf{y}}\mathsf{E}_{\mathbf{S}_1}[\mathrm{e}^{-it(c_1\mathbf{y}'\mathbf{S}_1^{-1}\mathbf{y}+c_2\log|\mathbf{S}_1|)}]\,\mathsf{E}_{\mathbf{S}_2}[\mathrm{e}^{it(c_1\mathbf{y}'\mathbf{S}_2^{-1}\mathbf{y}+c_2\log|\mathbf{S}_2|)}] \\
&= (n_1/n_2)^{itpc_2}\,\mathsf{E}_{\mathbf{y}}\mathsf{E}_{\mathbf{A}_1}[\mathrm{e}^{-it(c_1n_1\mathbf{y}'\mathbf{A}_1^{-1}\mathbf{y}+c_2\log|\mathbf{A}_1|)}] \\
&\quad\times \mathsf{E}_{\mathbf{A}_2}[\mathrm{e}^{it(c_1n_2\mathbf{y}'\mathbf{A}_2^{-1}\mathbf{y}+c_2\log|\mathbf{A}_2|)}].
\end{aligned} \tag{3.3}$$

Applying Lemma 2.7 to evaluate the expectations with respect to $\mathbf{A}_1$ and $\mathbf{A}_2$, we obtain

$$\begin{aligned}
\mathsf{E}[\mathrm{e}^{it\hat{Q}_2}] &= \left(\frac{n_1^p|\mathbf{\Sigma}_2|}{n_2^p|\mathbf{\Sigma}_1|}\right)^{itc_2} \frac{\pi^{p-1}\Gamma_{p-1}(\frac{1}{2}n_1 - itc_2)\,\Gamma_{p-1}(\frac{1}{2}n_2 + itc_2)}{\Gamma_p(\frac{1}{2}n_1)\,\Gamma_p(\frac{1}{2}n_2)} \\
&\quad\times \mathsf{E}_{\mathbf{y}}\left[\mathrm{B}^{(1)}_{itc_2-(N_1-p)/2}\left(\frac{1}{2}\,itc_1n_1\mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y}\right)\right. \\
&\quad\left.\times B^{(1)}_{-itc_2-(N_2-p)/2}\left(-\frac{1}{2}\,itc_1n_2\mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y}\right)\right].
\end{aligned} \tag{3.4}$$

Now suppose $\mathbf{y} \stackrel{d}{=} \mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}_1)$. Then by Lemma 2.8, eq. (2.19), we have

$$\begin{aligned}
&\mathsf{E}_{\mathbf{y}}[\mathrm{B}^{(1)}_{itc_2-(N_1-p)/2}(\tfrac{1}{2}itc_1n_1\mathbf{y}'\mathbf{\Sigma}_1^{-1}\mathbf{y})\,B^{(1)}_{-itc_2-(N_2-p)/2}(-\tfrac{1}{2}itc_1n_2\mathbf{y}'\mathbf{\Sigma}_2^{-1}\mathbf{y})] \\
&\quad= \int_0^\infty \int_0^\infty x_1^{itc_2}x_2^{-itc_2}\,\mathsf{E}[\mathrm{e}^{itc_1\sum_{j=1}^{p}(n_2x_2\lambda_j - n_1x_1)\mathrm{V}_j/2}] \\
&\quad\quad\times \prod_{g=1}^{2} \mathrm{e}^{-x_g^{-1}}x_g^{-(N_g-p)/2-1}\,dx_g.
\end{aligned} \tag{3.5}$$

Combining (3.4) and (3.5), we obtain the characteristic function of $\hat{Q}_2$ in the form

$$\mathsf{E}[e^{it\hat{Q}_2}] = \left(\frac{n_1^p|\boldsymbol{\Sigma}_2|}{n_2^p|\boldsymbol{\Sigma}_1|}\right)^{itc_2} \frac{\pi^{p-1}\Gamma_{p-1}(\frac{1}{2}n_1 - itc_2)\,\Gamma_{p-1}(\frac{1}{2}n_2 + itc_2)}{\Gamma_p(\frac{1}{2}n_1)\,\Gamma_p(\frac{1}{2}n_2)}$$

$$\times \int_0^\infty \int_0^\infty x_1^{itc_2} x_2^{-itc_2} \mathsf{E}[e^{itc_1 \sum_{j=1}^p (n_2 x_2 \lambda_j - n_1 x_1)\, \mathrm{V}_j/2}]$$

$$\times \prod_{g=1}^2 e^{-x_g^{-1}} x_g^{-(N_g - p)/2 - 1}\, dx_g. \tag{3.6}$$

Collecting all terms with an exponent itc_2 and applying Lemma 2.3, we find that (3.6) reduces to

$$\mathsf{E}[e^{it\hat{Q}_2}] = \frac{\pi^{p-1}\Gamma_{p-1}(\frac{1}{2}n_1)\,\Gamma_{p-1}(\frac{1}{2}n_2)}{\Gamma_p(\frac{1}{2}n_1)\,\Gamma_p(\frac{1}{2}n_2)}\, \mathsf{E}[e^{itc_2 \sum_{j=1}^{p-1} \log((N_2 - j)/(N_1 - j)\,\mathrm{F}_j)}]$$

$$\times \int_0^\infty \int_0^\infty \mathsf{E}[e^{itc_2 \log(n_1^p |\boldsymbol{\Sigma}_2|\, x_1 n_2^p |\boldsymbol{\Sigma}_1| x_2}]\, \mathsf{E}[e^{itc_1 \sum_{j=1}^p (n_2 x_2 \lambda_j - n_1 x_1)\, \mathrm{V}_j/2}]$$

$$\times \prod_{g=1}^2 e^{-x_g^{-1}} x_g^{-(N_g - p)/2 - 1}\, dx_g. \tag{3.7}$$

Simplifying the constant term in (3.7), we obtain

$$\mathsf{E}[e^{it\hat{Q}_2}]$$

$$= \int_0^\infty \int_0^\infty \mathsf{E}[e^{it[c_1 \sum_{j=1}^p (n_2 x_2 \lambda_j - n_1 x_1)\mathrm{V}_j/2 + c_2 \log(n_1^p x_1)/(n_2^p x_2)\, |\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2| + c_2 \sum_{j=1}^{p-1} \log((N_2 - j)/(N_1 - j)\,\mathrm{F}_j)]}]$$

$$\times \prod_{g=1}^2 \frac{x_g^{-(N_g - p)/2 - 1} e^{-x_g^{-1}}}{\Gamma(\frac{1}{2}(N_g - p))}\, dx_g. \tag{3.8}$$

On making the transformation $t_1 = 2x_1^{-1}$ and $t_2 = 2x_2^{-1}$, which has the Jacobian $4t_1^{-2}t_2^{-2}$, we obtain (3.8) in the form

$$\mathsf{E}[e^{it\hat{Q}_2}] = \int_0^\infty \int_0^\infty f_{N_1 - p}(t_1)\, f_{N_2 - p}(t_2)\, \mathsf{E}[e^{it\mathrm{W}(t_1,\, t_2)}]\, dt_1\, dt_2, \tag{3.9}$$

where $f_k(\cdot)$ denotes the density function of a chi-squared random variable with k degrees of freedom and

$$W(t_1, t_2) \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left(\frac{n_2}{t_2} \lambda_j - \frac{n_1}{t_1} \right) \mathrm{V}_j$$
$$+ c_2 \left[\log \left(\frac{n_1^p t_2}{n_2^p t_1} \right) - \log |\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1| + \sum_{j=1}^{p-1} \log \left(\frac{N_2 - j}{N_1 - j} \mathrm{F}_j \right) \right].$$

Interchanging expectation and integrals in (3.9), we obtain

$$\mathsf{E}[\mathrm{e}^{it\hat{Q}_2}] = \mathsf{E} \int_0^\infty \int_0^\infty f_{N_1 - p}(t_1) \, f_{N_2 - p}(t_2) \, \mathrm{e}^{it\mathrm{W}(t_1, t_2)} \, dt_1 \, dt_2,$$
$$\equiv \mathsf{E}[\mathrm{e}^{it\mathrm{W}(T_1, T_2)}]. \tag{3.10}$$

Therefore, from (3.10), we conclude that $\hat{Q}_2 \stackrel{d}{=} W$ where W is the random variable on the right-hand side of (3.1).

Finally, if $\mathbf{y} \in \mathbf{\Pi}_2$ then the same method of proof establishes (3.2). ∎

Substituting $c_2 = 0$ in (3.1) we obtain the following result.

COROLLARY 3.2. *Under the hypotheses of Theorem* 3.1, *the zero-means minimum-distance discriminant function* $\hat{Q}_4 = c_1 \mathbf{y}' [\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}] \, \mathbf{y}$ *is equal in distribution to*

$$c_1 \sum_{j=1}^{p} \left(\frac{n_2}{\mathrm{T}_2} \lambda_j - \frac{n_1}{\mathrm{T}_1} \right) \mathrm{V}_j, \tag{3.11}$$

if $\mathbf{y} \in \mathbf{\Pi}_1$, *and to*

$$c_1 \sum_{j=1}^{p} \left(\frac{n_2}{\mathrm{T}_2} - \frac{n_1}{\mathrm{T}_1} \frac{1}{\lambda_j} \right) \mathrm{V}_j, \tag{3.12}$$

if $\mathbf{y} \in \mathbf{\Pi}_2$.

In the case of $\hat{Q}_1 = c_1 (\mathbf{y} - \bar{\mathbf{y}})' \; [\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}] (\mathbf{y} - \bar{\mathbf{y}}) + c_2 \log |\mathbf{S}_2| / |\mathbf{S}_1|$, the equal-means plug-in discriminant function, we have the following result.

THEOREM 3.3. *Let* $\mathrm{T}_1, \mathrm{T}_2, \mathrm{V}_1, ..., \mathrm{V}_p$, *and* $\mathrm{F}_1, ..., \mathrm{F}_{p-1}$ *be mutually independent random variables, where* $\mathrm{T}_g \stackrel{d}{=} \chi^2_{N_g - p}$, $g = 1, 2$; $\mathrm{V}_j \stackrel{d}{=} \chi^2_1$, $j = 1, ..., p$;

and F_j *is F-distributed with* (N_2-j, N_1-j) *degrees of freedom,* $j=1, ..., p-1$. *Also, let* $m_g = N_g/(N_1+N_2)^2$, $g=1, 2$. *If* $\mathbf{y} \in \mathbf{\Pi}_1$ *then*

$$\hat{Q}_1 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left[\frac{n_2}{T_2} ((m_1+1)\lambda_j + m_2) - \frac{n_1}{T_1} (m_1 + 1 + m_2 \lambda_j^{-1}) \right] V_j$$
$$+ c_2 \left[\log \left(\frac{n_1^p T_2}{n_2^p T_1} \right) - \log |\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1| + \sum_{j=1}^{p-1} \log \left(\frac{N_2 - j}{N_1 - j} F_j \right) \right], \quad (3.13)$$

and if $\mathbf{y} \in \mathbf{\Pi}_2$ *then*

$$\hat{Q}_1 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left[\frac{n_2}{T_2} (m_1 \lambda_j + m_2 + 1) - \frac{n_1}{T_1} (m_1 + (m_2+1) \lambda_j^{-1}) \right] V_j$$
$$+ c_2 \left[\log \left(\frac{n_1^p T_2}{n_2^p T_1} \right) - \log |\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1| + \sum_{j=1}^{p-1} \log \left(\frac{N_2 - j}{N_1 - j} F_j \right) \right]. \quad (3.14)$$

Proof. Suppose $\mathbf{y} \in \mathbf{\Pi}_1$; then $\mathbf{y} - \bar{\mathbf{y}} \stackrel{d}{=} N_p(\mathbf{0}, (m_1+1)\mathbf{\Sigma}_1 + m_2 \mathbf{\Sigma}_2)$. By proceeding as in the proof of Theorem 3.1, with the aid of Lemma 2.9, we obtain (3.13). Also, the proof of (3.14) is obtained similarly. ∎

COROLLARY 3.4. *Under the hypotheses of Theorem* 3.3 *the equal-means minimum-distance plug-in discriminant function,* $\hat{Q}_3 = c_1 (\mathbf{y} - \bar{\mathbf{y}})' [\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}](\mathbf{y} - \bar{\mathbf{y}})$, *satisfies*

$$\hat{Q}_3 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left[\frac{n_2}{T_2} ((m_1+1)\lambda_j + m_2) - \frac{n_1}{T_1} \left(m_1 + 1 + m_2 \lambda_j^{-1} \right) \right] V_j \quad (3.15)$$

if $\mathbf{y} \in \mathbf{\Pi}_1$, *and*

$$\hat{Q}_3 \stackrel{d}{=} c_1 \sum_{j=1}^{p} \left[\frac{n_2}{T_2} (m_1 \lambda_j + m_2 + 1) - \frac{n_1}{T_1} (m_1 + (m_2+1) \lambda_j^{-1}) \right] V_j, \quad (3.16)$$

if $\mathbf{y} \in \mathbf{\Pi}_2$.

4. MISCLASSIFICATION PROBABILITIES

In this section we consider the behavior of the misclassification probabilities associated with the discriminant criteria $\hat{Q}_j$, $1 \leqslant j \leqslant 4$. Let $\hat{Q}$ denote any of these discriminant functions and k be the cut-off constant defined in (1.2). Then we use the notation $P(2 \mid 1) := P\{\hat{Q} \leqslant \log k \mid \mathbf{y} \in \mathbf{\Pi}_1\}$, the probability of misclassifying $\mathbf{y}$ to $\mathbf{\Pi}_2$ when, in fact, $\mathbf{y} \in \mathbf{\Pi}_1$; and $P(1 \mid 2) := P\{\hat{Q} > \log k \mid \mathbf{y} \in \mathbf{\Pi}_2\}$, the probability of misclassifying $\mathbf{y}$ to $\mathbf{\Pi}_1$ when, in fact, $\mathbf{y} \in \mathbf{\Pi}_2$.

In the following we retain the prior notation for the eigenvalues $\lambda_1, ..., \lambda_p$, and for the random variables T_1, T_2 and $V_1, ..., V_p$. We also denote by Y a random variable which is F-distributed with $(N_2 - p, N_1 - p)$ degrees of freedom, and which is independent of $V_1, ..., V_p$. We also assume that $q_1 = q_2$ and $C(1 \mid 2) = C(2 \mid 1)$, so that $\log k = 0$.

THEOREM 4.1. *The probabilities of misclassification for the equal-means discriminant function* $\hat{Q}_4 = c_1 \mathbf{y}'[\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}]\, \mathbf{y}$ *are given by*

$$P(2 \mid 1) = P\left\{\sum_{j=1}^{p} \left(\lambda_j - \frac{n_1(N_2 - p)}{n_2(N_1 - p)}\, \mathrm{Y}\right) \mathrm{V}_j \leqslant 0\right\} \tag{4.1}$$

and

$$P(1 \mid 2) = P\left\{\sum_{j=1}^{p} \left(\frac{n_2(N_1 - p)}{n_1(N_2 - p)}\, \frac{1}{\mathrm{Y}} - \frac{1}{\lambda_j}\right) \mathrm{V}_j > 0\right\}. \tag{4.2}$$

Further, $P(2 \mid 1)$ *is decreasing, and* $P(1 \mid 2)$ *is increasing, in* $(\lambda_1, ..., \lambda_p)$.

Proof. Since the proofs for $P(2 \mid 1)$ and $P(1 \mid 2)$ are similar, we provide details for $P(2 \mid 1)$ only.

By the stochastic representation (3.11) in Corollary 3.2, we obtain

$$\begin{aligned} P(2 \mid 1) &= \mathrm{P}\left\{c_1 \sum_{j=1}^{p} \left(\frac{n_2}{\mathrm{T}_2}\, \lambda_j - \frac{n_1}{\mathrm{T}_1}\right) \mathrm{V}_j \leqslant 0\right\} \\ &= \mathrm{P}\left\{\sum_{j=1}^{p} \left(\lambda_j - \frac{n_1 \mathrm{T}_2}{n_2 \mathrm{T}_1}\right) \mathrm{V}_j \leqslant 0\right\}. \end{aligned} \tag{4.3}$$

Since $\mathrm{T}_2/\mathrm{T}_1 \stackrel{d}{=} (N_2 - p)\, \mathrm{Y}/(N_1 - p)$ then the proof of (4.1) is complete.

Note also that the random variable in (4.3) is stochastically increasing in each λ_j; hence, $P(2 \mid 1)$ is decreasing in $(\lambda_1, ..., \lambda_p)$. ▌

In general, it does not seem likely that simple expressions can be obtained for the probabilities of misclassification. For special choices of p or $\lambda_1, ..., \lambda_p$, however, we have derived relatively simple reductions of (4.1) and (4.2). For example, if $\lambda_j = \lambda$ for all $j = 1, ..., p$, equivalently, $\mathbf{\Sigma}_1 = \lambda \mathbf{\Sigma}_2$, then (4.1) and (4.2) reduce to

$$P(2 \mid 1) = \mathrm{P}\left\{\mathrm{Y} \geqslant \frac{n_2(N_1 - p)}{n_1\,(N_2 - p)}\, \lambda\right\} \tag{4.4}$$

and

$$P(1 \mid 2) = \mathrm{P}\left\{\mathrm{Y} < \frac{n_2(N_1 - p)}{n_1(N_2 - p)}\, \lambda^{-1}\right\}, \tag{4.5}$$

respectively, where Y has an F-distribution with degrees of freedom $(N_2 - p, N_1 - p)$.

Let us now consider the case in which p is even, $p \geqslant 4$, $\lambda_1 = \cdots = \lambda_r \equiv \gamma_1$ and $\lambda_{r+1} = \cdots = \lambda_p \equiv \gamma_2$, where $1 \leqslant r < p$ and $\gamma_1 > \gamma_2$. To simplify the resulting exposition we adopt the notation $\delta := n_1(N_2 - p)/n_2(N_1 - p)$. Conditional on $\mathrm{Y} = y$, where $\gamma_2/\delta < y < \gamma_1/\delta$, we denote by $\mathrm{B}_{21}(y)$ and $\mathrm{B}_{12}(y)$ two binomial random variables on $\frac{1}{2}p - 1$ trials and probabilities of success $(\delta y - \gamma_2)/(\gamma_1 - \gamma_2)$ and $(\gamma_2^{-1} - \delta^{-1}y^{-1})/(\gamma_2^{-1} - \gamma_1^{-1})$, respectively. With these conventions, we have the following result.

COROLLARY 4.2. *For p even, $p \geqslant 4$, the probabilities of misclassification corresponding to the discriminant function $\hat{Q}_4$ are*

$$P(2 \mid 1) = P\left\{\mathrm{Y} \geqslant \frac{\gamma_1}{\delta}\right\} + \mathsf{E}_{\mathrm{Y}}\left[P\left\{\mathrm{B}_{21}(y) \geqslant \frac{r}{2}\right\} I\left(\frac{\gamma_2}{\delta} < y < \frac{\gamma_1}{\delta}\right) \middle| \mathrm{Y} = y \right], \tag{4.6}$$

and

$$P(1 \mid 2) = P\left\{\mathrm{Y} < \frac{\gamma_2}{\delta}\right\} + \mathsf{E}_{\mathrm{Y}}\left[P\left\{\mathrm{B}_{12}(y) \leqslant \frac{r}{2}\right\} I\left(\frac{\gamma_2}{\delta} < y < \frac{\gamma_1}{\delta}\right) \middle| \mathrm{Y} = y \right], \tag{4.7}$$

where $I(\cdot)$ denotes an indicator function. Moreover,

$$P\left\{\mathrm{Y} \geqslant \frac{\gamma_1}{\delta}\right\} \leqslant P(2 \mid 1) \leqslant P\left\{\mathrm{Y} \geqslant \frac{\gamma_2}{\delta}\right\} \tag{4.8}$$

and

$$P\left\{\mathrm{Y} \leqslant \frac{\gamma_2}{\delta}\right\} \leqslant P(1 \mid 2) \leqslant P\left\{\mathrm{Y} \leqslant \frac{\gamma_1}{\delta}\right\}. \tag{4.9}$$

Proof. By (4.1) we have

$$P(2 \mid 1) = \mathrm{P}\{(\gamma_1 - \delta \mathrm{Y})\, \mathrm{V}_1 + (\gamma_2 - \delta \mathrm{Y})\, \mathrm{V}_2 \leqslant 0\}, \tag{4.10}$$

where $\mathrm{V}_1 \overset{d}{=} \chi_r^2$, $\mathrm{V}_2 \overset{d}{=} \chi_{p-r}^2$, and Y, V_1, and V_2 are mutually independent. Define functions $h_1, h_2 \colon (0, \infty) \to \mathbb{R}$ by $h_g(y) = \gamma_g - \delta y$, $y > 0$, for $g = 1, 2$. Then (4.10) becomes

$$\begin{aligned} P(2 \mid 1) &= P\{h_1(\mathrm{Y})\, \mathrm{V}_1 + h_2(\mathrm{Y})\, \mathrm{V}_2 \leqslant 0\} \\ &= \mathsf{E}_{\mathrm{Y}} P\{h_1(y)\, \mathrm{V}_1 + h_2(y)\, \mathrm{V}_2 \leqslant 0 \mid \mathrm{Y} = y\}. \end{aligned} \tag{4.11}$$

Since $0<\gamma_2<\gamma_1$ then $h_2(y)<h_1(y)$ for all $y>0$; therefore

$$P\{h_1(y)\,\mathrm{V}_1+h_2(y)\,\mathrm{V}_2\leqslant 0\mid \mathrm{Y}=y\}=\begin{cases}0, & \text{if } \; h_2(y)\geqslant 0\\ 1, & \text{if } \; h_1(y)\leqslant 0.\end{cases} \tag{4.12}$$

For the case in which $h_2(y)<0<h_1(y)$, we have

$$\begin{aligned} P\{h_1(y)\,\mathrm{V}_1+h_2(y)\,\mathrm{V}_2\leqslant 0\mid \mathrm{Y}=y\} &= P\{h_1(y)\,\mathrm{V}_1\leqslant -h_2(y)\,\mathrm{V}_2\mid \mathrm{Y}=y\}\\ &= P\left\{\frac{\mathrm{V}_1}{\mathrm{V}_2}\leqslant -\frac{h_2(y)}{h_1(y)}\,\middle|\, \mathrm{Y}=y\right\}\\ &= P\left\{\mathrm{X}\leqslant -\frac{(p-r)}{r}\,\frac{h_2(y)}{h_1(y)}\,\middle|\, \mathrm{Y}=y\right\}, \end{aligned} \tag{4.13}$$

where $\mathrm{X}=(p-r)\,\mathrm{V}_1/r\mathrm{V}_2$ has an F-distribution with $(r,\,p-r)$ degrees of freedom.

Since p is even and $p\geqslant 4$ then, by a well-known result (cf. Johnson and Kotz [7, p. 88]),

$$P\left\{\mathrm{X}\leqslant\frac{(p-r)}{r}x\right\}=P\left\{\mathrm{B}\geqslant\frac{1}{2}r\right\},\qquad x>0, \tag{4.14}$$

where B has a binomial distribution with $\frac{1}{2}p-1$ trials and probability of success $x/(x+1)$. Therefore, by (4.12), (4.13), and (4.14), we have

$$\begin{aligned} &\mathrm{P}\{h_1(y)\,\mathrm{V}_1+h_2(y)\,\mathrm{V}_2\leqslant 0\mid \mathrm{Y}=y\}\\ &\qquad=\begin{cases}0, & \text{if } \; y\leqslant \gamma_2/\delta\\ \mathrm{P}\{\mathrm{B}_{21}(y)\geqslant\frac{1}{2}r\}, & \text{if } \; \gamma_2/\delta<y<\gamma_1/\delta\\ 1, & \text{if } \; y\geqslant\gamma_1/\delta.\end{cases} \end{aligned} \tag{4.15}$$

By (4.11) and (4.15), and the law of total probability, we obtain

$$\begin{aligned} P(2\mid 1)&=\mathsf{E}_{\mathrm{Y}}\left[I\left\{y\geqslant\frac{\gamma_1}{\delta}\right\}\,\middle|\,\mathrm{Y}=y\right]\\ &\quad+\mathsf{E}_{\mathrm{Y}}\left[\mathrm{P}\left\{\mathrm{B}_{21}(y)\geqslant\frac{r}{2}\right\}I\left(\frac{\gamma_2}{\delta}<y<\frac{\gamma_1}{\delta}\right)\,\middle|\,\mathrm{Y}=y\right]; \end{aligned} \tag{4.16}$$

hence (4.6) follows from (4.16).

In the case of $P(1 \mid 2)$, we define $\tilde{h}_g(y) = \delta^{-1} y^{-1} - \gamma_g^{-1}$, $y > 0$, for $g = 1, 2$. Then

$$P(1 \mid 2) = \mathsf{E}_{\mathrm{Y}} \left[\mathrm{P}\{ \tilde{h}_1(y)\, \mathrm{V}_1 + \tilde{h}_2(y)\, \mathrm{V}_2 > 0 \mid \mathrm{Y} = y \} \right].$$

Proceeding similarly as in the remainder of the calculation of $P(2 \mid 1)$, we obtain (4.7).

By applying to (4.6) the inequality $\mathrm{P}\{ \mathrm{B}_{21}(y) \geqslant r/2 \} \geqslant 0$ we obtain

$$P(2 \mid 1) \geqslant P\left\{ \mathrm{Y} \geqslant \frac{\gamma_1}{\delta} \right\},$$

which is the lower bound in (4.8). Further, by applying to (4.6) the inequality $\mathrm{P}\{ \mathrm{B}_{21}(y) \geqslant r/2 \} \leqslant 1$ we obtain

$$\begin{aligned} P(2 \mid 1) &\leqslant \mathrm{P}\left\{ \mathrm{Y} \geqslant \frac{\gamma_1}{\delta} \right\} + \mathsf{E}_{\mathrm{Y}} \left[I\left(\frac{\gamma_2}{\delta} < y < \frac{\gamma_1}{\delta} \right) \Big| \, \mathrm{Y} = y \right] \\ &= \mathrm{P}\left\{ \mathrm{Y} \geqslant \frac{\gamma_1}{\delta} \right\} + \mathrm{P}\left\{ \frac{\gamma_2}{\delta} < Y < \frac{\gamma_1}{\delta} \right\} \\ &= \mathrm{P}\left\{ \mathrm{Y} \geqslant \frac{\gamma_2}{\delta} \right\}, \end{aligned}$$

which is the upper bound in (4.8).

A similar argument applies to establish (4.9), and then the proof is complete. ∎

For the case in which $p = 2$, Theorem 4.1 reduces to the following result.

COROLLARY 4.3. *For $p = 2$, the misclassification probabilities for $\hat{Q}_4$ are*

$$P(2 \mid 1) = P\left\{ \mathrm{Y} \geqslant \frac{\lambda_2}{\delta} \right\} - \frac{2}{\pi} \mathsf{E}_{\mathrm{Y}} \left[I\left(\frac{\lambda_2}{\delta} < y < \frac{\lambda_1}{\delta} \right) \sin^{-1} \sqrt{\frac{\lambda_1 - \delta y}{\lambda_1 - \lambda_2}} \, \Bigg| \, \mathrm{Y} = y \right], \tag{4.17}$$

and

$$P(1 \mid 2) = P\left\{ \mathrm{Y} < \frac{\lambda_1}{\delta} \right\} - \frac{2}{\pi} \mathsf{E}_{\mathrm{Y}} \left[I\left(\frac{\lambda_2}{\delta} < y < \frac{\lambda_1}{\delta} \right) \sin^{-1} \sqrt{\frac{\lambda_2^{-1} - \delta^{-1} y^{-1}}{\lambda_2^{-1} - \lambda_1^{-1}}} \, \Bigg| \, \mathrm{Y} = y \right]. \tag{4.18}$$

Proof. For $p=2$ and $r=1$, it follows by (4.12) and (4.13) that

$$\begin{aligned} \mathrm{P}\{h_1(y)\,\mathrm{V}_1+h_2(y)\,\mathrm{V}_2\leqslant 0\mid \mathrm{Y}=y\} &= \mathrm{P}\{h_1(y)\,\mathrm{V}_1\leqslant -h_2(y)\,\mathrm{V}_2\mid \mathrm{Y}=y\} \\ &= \mathrm{P}\left\{\frac{\mathrm{V}_1}{\mathrm{V}_2}\leqslant -\frac{h_2(y)}{h_1(y)}\,\middle|\, \mathrm{Y}=y\right\} \\ &= \mathrm{P}\left\{\mathrm{X}\leqslant -\frac{h_2(y)}{h_1(y)}\,\middle|\, \mathrm{Y}=y\right\}, \end{aligned} \tag{4.19}$$

where X has an F-distribution with (1, 1) degrees of freedom. Using the well-known relationship between the F- and beta distributions (cf. Johnson and Kotz [7, p. 78]), we obtain

$$\begin{aligned} \mathrm{P}\left\{\mathrm{X}\leqslant -\frac{h_2(y)}{h_1(y)}\,\middle|\, \mathrm{Y}=y\right\} &= 1-\mathrm{P}\left\{\mathrm{U}\leqslant \frac{\lambda_1-\delta y}{\lambda_1-\lambda_2}\,\middle|\, \mathrm{Y}=y\right\} \\ &= 1-\frac{2}{\pi}\sin^{-1}\sqrt{\frac{\lambda_1-\delta y}{\lambda_1-\lambda_2}} \end{aligned} \tag{4.20}$$

if $h_2(y)<0<h_1(y)$, where U has a beta distribution with parameters $(\frac{1}{2}, \frac{1}{2})$. Therefore, applying the same reasoning as in (4.12) and (4.13), it follows from (4.20) that

$$\mathrm{P}\{h_1\mathrm{V}_1+h_2\mathrm{V}_2\leqslant 0\mid \mathrm{Y}=y\}$$

$$= \begin{cases} 0, & \text{if } \; y\leqslant \lambda_2/\delta \\ 1-\dfrac{2}{\pi}\sin^{-1}\sqrt{\dfrac{\lambda_1-\delta y}{\lambda_1-\lambda_2}}, & \text{if } \; \lambda_2/\delta<y<\lambda_1/\delta \\ 1, & \text{if } \; y\geqslant \lambda_1/\delta. \end{cases} \tag{4.21}$$

By (4.21) and the law of total probability

$$\begin{aligned} P(2\mid 1) &= \mathsf{E}_{\mathrm{Y}}\left[I\left\{y\geqslant \frac{\lambda_1}{\delta}\right\}\,\middle|\, \mathrm{Y}=y\right] \\ &\quad + \mathsf{E}_{\mathrm{Y}}\left[\left(1-\frac{2}{\pi}\sin^{-1}\sqrt{\frac{\lambda_1-\delta y}{\lambda_1-\lambda_2}}\right) I\left(\frac{\lambda_2}{\delta}<y<\frac{\lambda_1}{\delta}\right)\,\middle|\, \mathrm{Y}=y\right], \end{aligned}$$

which reduces to (4.17).

Finally, the proof of (4.18) is similar. ∎

In the case of the discriminant function $\hat{Q}_3 = c_1(\mathbf{y}-\bar{\mathbf{y}})' \, [\mathbf{S}_2^{-1}-\mathbf{S}_1^{-1}] \, (\mathbf{y}-\bar{\mathbf{y}})$, it follows from Corollary 3.4 that the associated probabilities of misclassification are

$$\begin{aligned} P(2 \mid 1) &= \mathrm{P}\left\{c_1 \sum_{j=1}^{p} \left[\frac{n_2}{\mathrm{T}_2}\left((m_1+1)\,\lambda_j + m_2\right) - \frac{n_1}{\mathrm{T}_1}\left(m_1+1+\frac{m_2}{\lambda_j}\right)\right] \mathrm{V}_j \leqslant 0\right\} \\ &= \mathrm{P}\left\{\frac{n_2 \mathrm{T}_1}{n_1 \mathrm{T}_2} \, \frac{\sum_{j=1}^{p} ((m_1+1)\,\lambda_j + m_2)\,\mathrm{V}_j}{\sum_{j=1}^{p} (m_1+1+m_2\lambda_j^{-1})\,\mathrm{V}_j} \leqslant 1\right\} \end{aligned} \tag{4.22}$$

and

$$\begin{aligned} P(1 \mid 2) &= \mathrm{P}\left\{c_1 \sum_{j=1}^{p} \left[\frac{n_2}{\mathrm{T}_2}\,(m_1\lambda_j + m_2 + 1) - \frac{n_1}{\mathrm{T}_1}\left(m_1 + \frac{(m_2+1)}{\lambda_j}\right)\right] \mathrm{V}_j > 0\right\} \\ &= \mathrm{P}\left\{\frac{n_2 \mathrm{T}_1}{n_1 \mathrm{T}_2} \, \frac{\sum_{j=1}^{p} (m_1\lambda_j + m_2 + 1))\,\mathrm{V}_j}{\sum_{j=1}^{p} (m_1 + (m_2+1)\,\lambda_j^{-1})\,\mathrm{V}_j} > 1\right\}. \end{aligned} \tag{4.23}$$

It is clear that the random variables in (4.22) and (4.23) are stochastically increasing in each λ_j. Therefore it follows that $P(2 \mid 1)$ is decreasing, and $P(1 \mid 2)$ is increasing, in $(\lambda_1, ..., \lambda_p)$.

In the case of the discriminant functions $\hat{Q}_1$ and $\hat{Q}_2$, it is more difficult to ascertain the behavior of $P(2 \mid 1)$ and $P(1 \mid 2)$ as functions of $\lambda_1, ..., \lambda_p$. The obstacle to complete analyses is the presence of the log -function in the stochastic representations given in Theorems 3.1 and 3.3. We can provide only the following brief comments in the case of $\hat{Q}_1$ (the case of $\hat{Q}_2$ is similar). Suppose that, say, λ_1 is allowed to vary and $\lambda_2, ..., \lambda_p$ are fixed; and assume $N_2 > p+4$. It follows from (3.13) that, for $\mathbf{y} \in \Pi_1$,

$$\mathsf{E}(\hat{Q}_1) = \sum_{j=1}^{p} (a_1\lambda_j - a_2\lambda_j^{-1} - c_2 \log \lambda_j) + a_3, \tag{4.24}$$

where a_1, a_2 and a_3 are positive constants; for later purposes we also note that

$$a_1 = c_1 n_2 (m_1+1)\,\mathsf{E}(V_1/T_2) = \frac{c_1 n_2 (m_1+1)}{N_2 - p - 2}.$$

It follows from (4.24) that, for sufficiently large λ_1, $\mathsf{E}(\hat{Q}_1)$ is strictly increasing in λ_1.

This suggests the possibility that, for sufficiently large λ_1, $P(2 \mid 1) = \mathrm{P}(\hat{Q}_1 > 0 \mid \mathbf{y} \in \Pi_1)$ is increasing in λ_1; however, there is conflicting evidence on this point. Indeed, letting $\lambda_1 \to \infty$ it follows from (4.24) that $\mathsf{E}(\hat{Q}_1) \sim a_1 \lambda_1$. Further we deduce from the stochastic representation (3.13) for $\hat{Q}_1$ that

$$\begin{aligned} \operatorname{Var}(\hat{Q}_1) &\sim c_1^2 n_2^2 (m_1+1)^2 \lambda_1^2 \operatorname{Var}(V_1/T_2) \\ &= c_1^2 n_2^2 (m_1+1)^2 \lambda_1^2 \left[\frac{2}{(N_2-p-2)(N_2-p-4)} - \frac{1}{(N_2-p-2)^2} \right]. \end{aligned}$$

Hence

$$\begin{aligned} \lim_{\lambda_1 \to \infty} \frac{\operatorname{Var}(\hat{Q}_1)}{[\mathsf{E}(\hat{Q}_1)]^2} &= (N_2-p-2)^2 \left[\frac{2}{(N_2-p-2)(N_2-p-4)} - \frac{1}{(N_2-p-2)^2} \right] \\ &= 1 + \frac{4}{N_2-p-4} > 1. \end{aligned}$$

Thus, although $\mathsf{E}(\hat{Q}_1)$ eventually increases as $\lambda_1 \to \infty$, the coefficient of variation of $\hat{Q}_1$ converges to a number greater than 1; this indicates a correspondingly high variability in the values of $\hat{Q}_1$ for large λ_1.

5. APPLICATIONS

In this section we apply the results of Section 4 to estimate the exact probabilities of misclassification for some data sets drawn from the literature on discriminant analysis. Each data set is introduced by means of a biplot (cf. Gabriel [5], Khattree and Naik [9]), the well-known two-dimensional graphical tool for displaying relationships among the individual observations, variables and population groups in a multivariate data set. In this plot, individual observations are represented by points while variables are represented by two-dimensional vectors. As recommended by Khattree and Naik [9, pp. 28-29], we use coordinates for the biplot which correspond to a principal components representation of the data; this also provides us with a measure of the proportion of total sample variability explained by the biplot. We shall also use the biplot as a dimension-reduction tool for eliminating highly correlated variables, and for illustrating the relative proximities of the means of the training samples.

Using the observations coordinate on the biplot, we construct 95% confidence ellipses for each population (cf. Johnson and Wichern [8, p. 189]). These ellipses illustrate graphically the covariance structure (i.e. the shape, volume, and orientation of the sample covariance matrices) of each population, and therefore provide us with a graphical interpretation for the results of our simulations.

Next, we apply Mardia's statistic [11] to test the hypothesis that each training sample is drawn from a normal population. Once we fail to reject the null hypothesis of normality, we select one or more of the discriminant functions $\hat{Q}_j$ and estimate the probabilities of misclassification. In all instances, we obtain our estimates by performing 100,000 iterations of Monte Carlo simulations of the stochastic representations in Section 4.

Throughout this section we will assume equal prior probabilities and equal costs of misclassification, so that $c_1 = c_2 = 1/2$; although these assumptions are not essential to our calculations, they are necessary for us to make proper comparisons between our estimates and previously published estimates of the probabilities of misclassification for Stocks' data. Note also that under these assumptions, the total probability of misclassification (TPMC) is

$$\mathrm{TPMC} = \tfrac{1}{2} P(1 \mid 2) + \tfrac{1}{2} P(2 \mid 1).$$

5.1. *Studentized Stochastic Representations*

The stochastic representations in Section 4 depend upon the unknown parameters $\lambda_1, ..., \lambda_p$ and $|\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1|$. Therefore, in estimating misclassification probabilities for a given data set we shall replace each of these parameters with its maximum likelihood estimate.

Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be the matrices of squares and products given in (2.1), $l_1, ..., l_p$ be the eigenvalues of $\mathbf{A}_2^{-1} \mathbf{A}_1$, and assume that $\lambda_1, ..., \lambda_p$ are distinct with $\lambda_1 > \cdots > \lambda_p$. By Muirhead [14, p. 20], the maximum likelihood estimate, $\hat{\lambda}_j$, of λ_j is of the form $\hat{\lambda}_j = \tilde{\lambda}_j + O((n_1 + n_2)^{-2})$ where

$$\tilde{\lambda}_j = \frac{n_2 - p - 1}{n_1}\, l_j - \frac{n_1 + n_2}{n_1^2}\, l_j \sum_{\substack{k=1 \\ k \neq j}}^{p} \frac{l_k}{l_j - l_k}, \qquad j = 1, ..., p. \tag{5.1}$$

Further, the estimators $\hat{\lambda}_j$ are asymptotically unbiased; in fact, $\mathsf{E}(\hat{\lambda}_j) = \lambda_j + O((n_1 + n_2)^{-2})$, $j = 1, ..., p$.

Denote by $\psi(\cdot)$ the digamma function, $\psi(x) = [\log \Gamma(x)]'$, $x > 0$; then it is well-known (cf. McLachlan [13, p. 57, eq. (3.2.8)]) that

$$\Psi(\mathbf{A}_1, \mathbf{A}_2) = \log |\mathbf{A}_2^{-1} \mathbf{A}_1| - \sum_{j=1}^{p} \left[\psi\left(\frac{N_1 - j}{2}\right) - \psi\left(\frac{N_2 - j}{2}\right) \right] \tag{5.2}$$

is an unbiased estimator of $\log |\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1|$. Note also that in the cases in which $N_1 = N_2$, this estimator reduces to $\log |\mathbf{A}_2^{-1} \mathbf{A}_1|$.

In order to estimate the misclassification probabilities for a given discriminant function, we replace the unknown parameters with the estimators given in (5.1) and (5.2). For example, in the case of the discriminant function $\hat{Q}_2$, for which the stochastic representation is given in Theorem 3.1, we estimate $P(2 \mid 1)$ and $P(1 \mid 2)$ through Monte Carlo simulation of the Studentized version,

$$\tilde{Q}_2 = c_1 \sum_{j=1}^{p} \left[\frac{n_2}{\mathrm{T}_2} \tilde{\lambda}_j - \frac{n_1}{\mathrm{T}_1} \right] \mathrm{V}_j + c_2 \left[\log \left(\frac{n_1^p \mathrm{T}_2}{n_2^p \mathrm{T}_1} \right) - \Psi(\mathbf{A}_1, \mathbf{A}_2) + \sum_{j=1}^{p-1} \log \left(\frac{N_2 - j}{N_1 - j} \mathrm{F}_j \right) \right], \tag{5.3}$$

where T_1, T_2, $\mathrm{V}_1, ..., \mathrm{V}_p$ and $\mathrm{F}_1, ..., \mathrm{F}_{p-1}$ are the random variables specified earlier.

We remark also that SAS macros, based on PROC IML, for estimating the probabilities of misclassification for any given data set using the results in this paper, are available from the authors.

5.2. *Stocks' Twins Data*

In a study of the physical characteristics of twins, Stocks [19] reported measurements of 14 variables made on 440 girls and 392 boys in Elementary and Central schools of the London County Council during the period 1925 to 1927. Of these 832 children, 563 were members of twin pairs, including 106 like-sexed pairs. Of the 563 twins, Stocks used an empirical criterion, based on fingerprints, height and head measurements, to classify the like-sexed twins into monozygotic or dizygotic population groups.

Okamoto [16] also studied the distribution theory of discriminant functions under the assumption of equal population means. He gave an application to the problem of classifying twins into monozygotic and dizygotic populations, based on a set of ten anthropological characteristics taken on a large sample of like-sexed twins from Osaka City, Japan. Unfortunately, the raw data collected during that study were not provided in [16], so we are unable to estimate the corresponding misclassification probabilities. (Although summary statistics were provided, they appear to contain some typographical errors; in fact, the sample covariance matrix provided for the training sample of dizygotic twins has a negative eigenvalue.)

Subsequent to the appearance of [16], several articles (cf. Bartlett and Please [2], Desu and Geisser [4], Young *et al.* [21], and Marco *et al.* [10]) provided additional analyses of Stocks' data. These authors chose random samples of 30 monozygotic and 30 dizygotic twins such that there were 15 female twins and 15 male twins in each group, based on the following subset of 10 variables from Stocks' original 14 variables:

$Y_1 = \text{Height}$	$Y_6 = \text{Interpupillary distance}$
$Y_2 = \text{Weight}$	$Y_7 = \text{Systolic blood pressure}$
$Y_3 = \text{Head length}$	$Y_8 = \text{Pulse interval}$
$Y_4 = \text{Head breadth}$	$Y_9 = \text{Strength of grip in the right hand}$
$Y_5 = \text{Head circumference}$	$Y_{10} = \text{Strength of grip in the left hand}$

As in [16], for each group of twins the difference between the first and second twin is taken as the observation, and the common mean of each population is assumed to be the zero vector. The biplot in Fig. 1 displays the original data collected by Stocks [19], except for incomplete observations; this totals 46 twin-pairs of each group.

We observe that the vectors corresponding to the variables Y_1, Y_9 and Y_{10} are nearly collinear, indicating high correlations between these variables. Since the vector corresponding to Y_{10} is the longest among these three vectors, indicating that Y_{10} has largest variability, we discarded Y_1 and Y_9. For similar reasons, we also discarded the variable Y_6. The ensuing modified biplot is displayed in Fig. 2.

It is noticeable that in the biplot in Fig. 2, the vectors corresponding to the variables Y_3 and Y_5 are relatively close; this suggests that we could have discarded Y_3, thereby lowering the dimension still further. Nevertheless we choose to retain Y_3, retaining more data about the populations.

In Table I are the sample means and covariance matrices for the training samples from $\mathbf{\Pi}_1$ (the monozygotic population) and $\mathbf{\Pi}_2$ (the dizygotic population).

We applied Mardia's test for multivariate normality to Stocks' data. The smallest P-value for Mardia's tests for skewness and kurtosis, for either $\mathbf{\Pi}_1$ or $\mathbf{\Pi}_2$, was 0.1078. Having failed to reject the hypothesis of normality we proceed to estimate the probabilities of misclassification.

Using the estimated discriminant function $\tilde{Q}_2$ given in (5.3), we performed a Monte Carlo simulation to estimate the probabilities of misclassification. The estimates obtained for $P(1 \mid 2)$, $P(2 \mid 1)$ and TPMC were 0.2321, 0.2098 and 0.2209, respectively.

It is timely to compare these estimated probabilities of misclassification with other published estimates. Bartlett and Please [2], under the assumption of the *uniform covariance structure* given in (1.12) where all parameters are assumed to be known, estimated an upper bound for the TPMC; Marco *et al.* [10], working with the same uniform covariance structure assumption, but with all the parameters unknown, obtained an asymptotic estimate of the TPMC; Young *et al.* [21], with no covariance restrictions but with all parameters are assumed known, estimated an upper bound for the TPMC; all these authors utilized an analog of $\hat{Q}_2$. We also applied the asymptotic approximations of Okamoto [16] to estimate the misclassification probabilities; since Okamoto's approximations involve the largest eigenvalue of $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1$, we used the estimator $\tilde{\lambda}_1$ in (5.1) to "Studentize" his approximation.

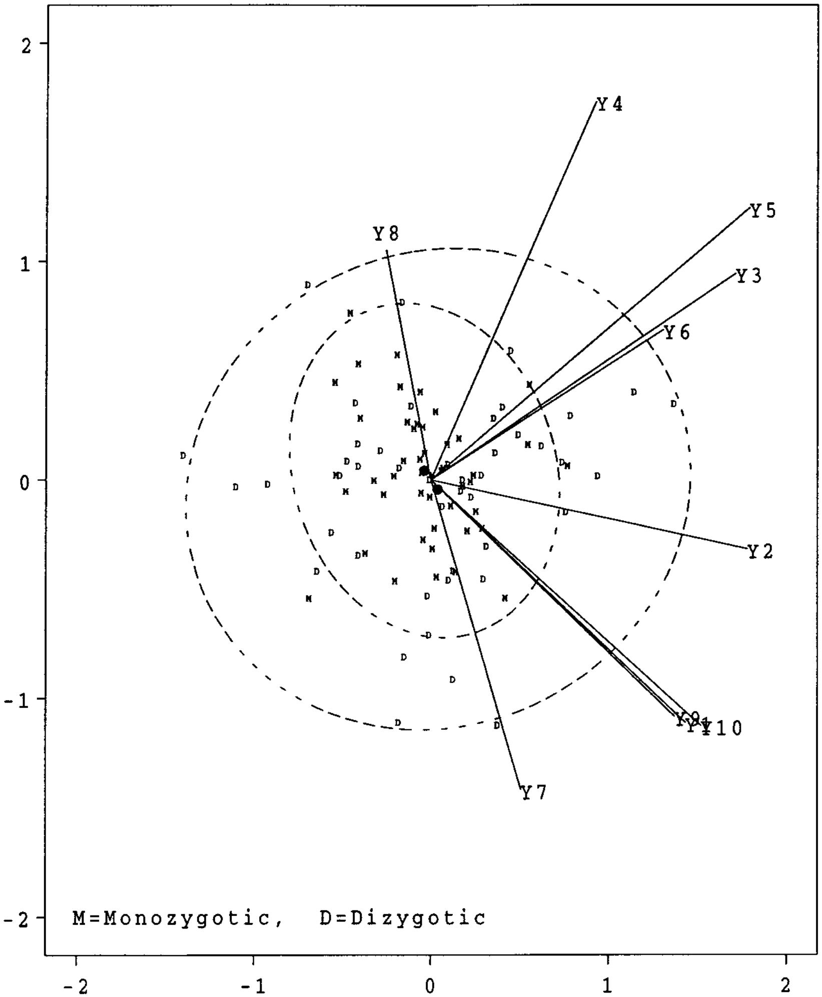

FIG. 1. Biplot of Stocks' data; estimated total variability explained $=55.40\%$.

Recall that we modified Stocks' data set, reducing the number of variables from ten to seven, because the biplot in Fig. 2 indicated a high correlation among several variables. Therefore our estimates of the misclassification probabilities are not directly comparable to those given by the aforementioned authors. We also applied Mardia's test for multivariate normality to this 10-variable data set; the null hypothesis of normality was rejected, with P-values of 0.0010 and 0.0437 for the skewness statistics for $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, respectively.

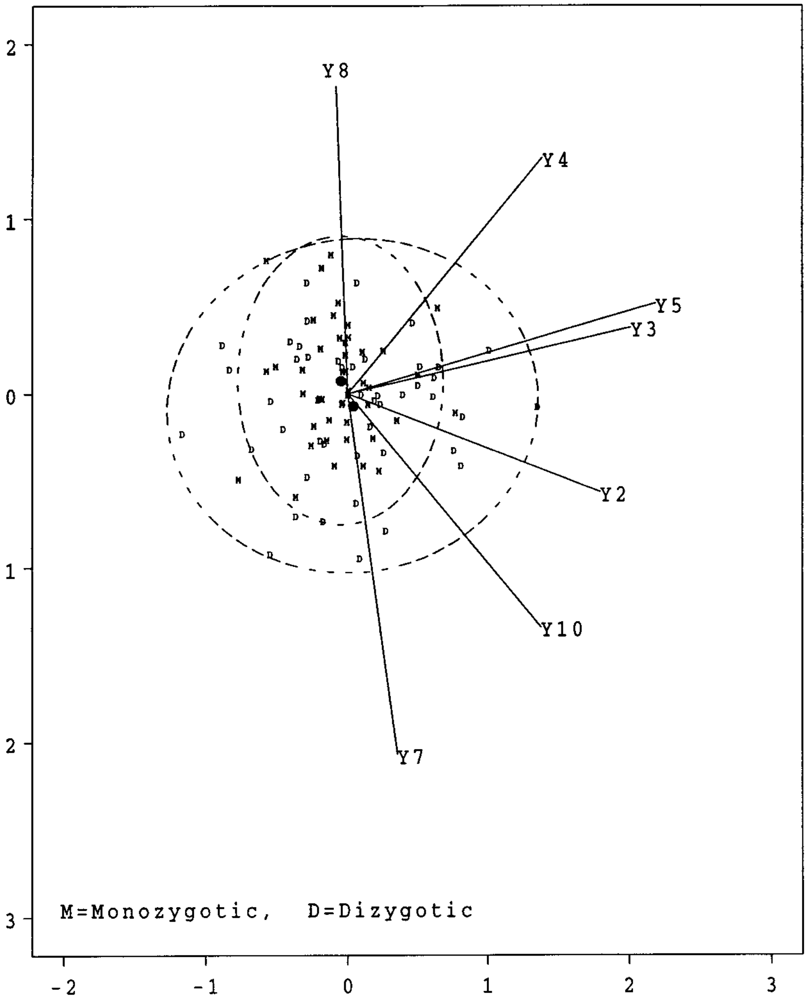

FIG. 2. Biplot of Stocks' reduced data; estimated total variability explained $=61.71\%$.

TABLE I

Summary Statistics for Stocks' Reduced Data

Sample means	Y_2	Y_3	Y_4	Y_5	Y_7	Y_8	Y_{10}
Monozygotic ($\mathbf{\Pi}_1$)	−1.9348	−0.3478	0.4565	−1.0217	−2.6304	0.4130	−0.6304
Dizygotic ($\mathbf{\Pi}_2$)	0.0217	0.8261	−0.2391	1.1739	1.8913	−1.6304	−0.4565

Monozygotic ($\mathbf{\Pi}_1$)	Y_2	Y_3	Y_4	Y_5	Y_7	Y_8	Y_{10}
Y_2	402.9957	27.7565	24.7251	92.2459	15.4865	33.7058	15.8643
Y_3	27.7565	9.3430	0.6512	15.3923	−5.1575	3.9691	1.2425
Y_4	24.7251	0.6512	8.9203	9.8324	−3.7280	5.3628	0.4053
Y_5	92.2459	15.3923	9.8324	52.8662	1.7415	8.6092	3.7638
Y_7	15.4865	−5.1575	−3.7280	1.7415	111.5271	−32.5338	3.7048
Y_8	33.7058	3.9691	5.3628	8.6092	−32.5338	65.5812	1.9551
Y_{10}	15.8643	1.2425	0.4053	3.7638	3.7048	1.9551	6.7271

Dizygotic ($\mathbf{\Pi}_2$)	Y_2	Y_3	Y_4	Y_5	Y_7	Y_8	Y_{10}
Y_2	821.8440	78.7594	22.5831	205.8850	78.6469	−26.8749	45.0546
Y_3	78.7594	28.5024	12.3130	66.3420	−0.4860	−4.8676	5.4522
Y_4	22.5831	12.3130	22.8082	47.9536	−4.9377	1.6903	1.0217
Y_5	205.8850	66.3420	47.9536	208.0135	13.0860	−2.7101	13.5923
Y_7	78.6469	−0.4860	−4.9377	13.0860	102.4546	−18.9812	6.5048
Y_8	−26.8749	−4.8676	1.6903	−2.7101	−18.9812	116.4604	−3.1609
Y_{10}	45.0546	5.4522	1.0217	13.5923	6.5048	−3.1609	9.3203

Nevertheless, we performed Monte Carlo simulations using all ten variables and the estimated discriminant functions $\tilde{Q}_j$, $j = 1, ..., 4$. Table II summarizes all available estimates of the misclassification probabilities for Stocks' ten-variable data set.

TABLE II

Estimated Probabilities of Misclassification for Stocks' Ten-Variable Data Set

Method	$N_1 = N_2$	$P(1 \mid 2)$	$P(2 \mid 1)$	TPMC	Upper bound
Bartlett *et al.*	30	N/A	N/A	N/A	0.3166
Marco *et al.*	30	N/A	N/A	0.1369 ± 0.0316	N/A
Young *et al.*	30	N/A	N/A	N/A	0.2994
Okamoto	46	0.2431	0.4052	0.3242	N/A
$\hat{Q}_1$	46	0.1173	0.1917	0.1546	N/A
$\hat{Q}_2$	46	0.1204	0.1831	0.1518	N/A
$\hat{Q}_3$	46	0.0139	0.7071	0.3605	N/A
$\hat{Q}_4$	46	0.0129	0.6959	0.3544	N/A

In Table II we see that the minimum distance discriminant functions $\hat{Q}_3$ and $\hat{Q}_4$ provide the highest estimates of $P(2 \mid 1)$, the probability of misclassifying an observation which belongs to the monozygotic population. To explain this phenomenon, observe that $\hat{Q}_3$ and $\hat{Q}_4$ depend only on the statistical distances between $\mathbf{y}$ and the pooled mean. Moreover, it is noticeable from Fig. 1 that the ellipse corresponding to the monozygotic group is almost entirely a subset of the ellipse corresponding to the dizygotic group. Therefore it is relatively easy to misclassify a monozygotic twin as dizygotic using the criteria $\hat{Q}_3$ or $\hat{Q}_4$; this explains the correspondingly high estimates of $P(2 \mid 1)$.

Conversely, the biplot reveals observations from the dizygotic group which are far outside the range of the monozygotic ellipse, so it should be more difficult to misclassify a dizygotic twin as monozygotic. Not surprisingly, the estimates of $P(1 \mid 2)$ derived from $\hat{Q}_3$ and $\hat{Q}_4$ are relatively small.

The discriminant functions $\hat{Q}_1$ and $\hat{Q}_2$ are dependent not only on the statistical distances from the pooled mean, but also on $|\mathbf{S}_1|$ and $|\mathbf{S}_2|$; thus $\hat{Q}_1$ and $\hat{Q}_2$ reflect differences in the generalized variances. From the biplot in Fig. 1, we see that the ellipses are very different in shape, volume, and orientation. Therefore, it is to be expected that $\hat{Q}_1$ and $\hat{Q}_2$ will provide more accurate estimates of the misclassification probabilities than $\hat{Q}_3$ or $\hat{Q}_4$.

5.3. *Rencher's Data on the Heads of Football Players*

Rencher [17, pp. 306–307], provided measurements on the heads of football players collected during a study on possible link between football helmet design and neck injuries. There were 30 subjects in each of three groups: high school football players (group 1), college football players (group 2), and non-football players (group 3). Six head measurements were made on each subject:

1. Head width at widest dimension (WDIM)
2. Head circumference (CIRCUM)
3. Front-to-back measurement at eye level (FBEYE)
4. Eye-to-top-of-head measurement (EYEHD)
5. Ear-to-top-of-head measurement (EARHD)
6. Jaw width (JAW)

The biplot in Fig. 3 indicates the samples of college football players and non-football players to be closely clustered together; moreover, the corresponding sample means are located near to each other. We note also the great similarity in location and shape of the two 95% confidence ellipses; this suggests that the population covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are not greatly dissimilar. Therefore it is reasonable to expect that our estimates of $P(1 \mid 2)$ and $P(2 \mid 1)$ will be relatively close.

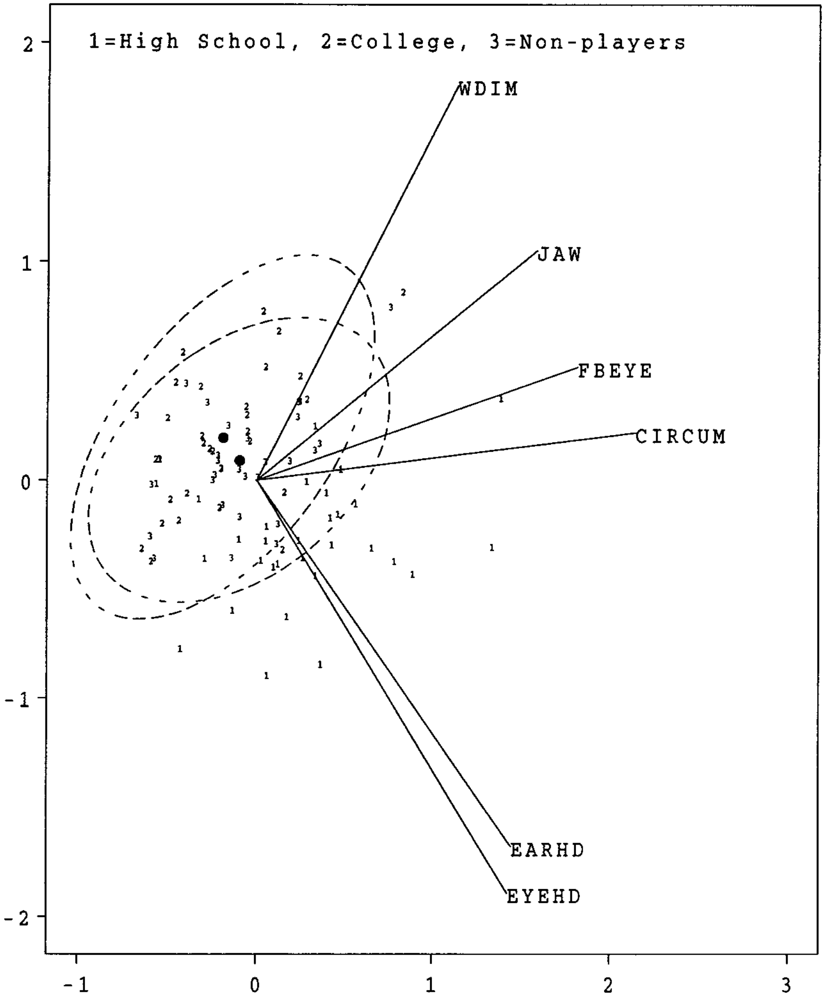

FIG. 3. Biplot of Rencher's football data; estimated total variability explained =68.69%.

We will assume that the populations of college players ($\mathbf{\Pi}_1$) and non-players ($\mathbf{\Pi}_2$) have equal means, and proceed to estimate the corresponding misclassification probabilities.

The sample means and sample covariance matrices for the college players and non-football players are given in Table III.

We applied Mardia's test to the head size data; the smallest P-value obtained for Mardia's tests for skewness and kurtosis, for either $\mathbf{\Pi}_1$ or $\mathbf{\Pi}_2$, was 0.2909. Therefore we fail to reject the null hypothesis of multivariate normality.

TABLE III

Summary Statistics for Football Players' Head Size Data

Sample means	WDIM	CIRCUM	FBEYE	EYEHD	EARHD	JAW
College players ($\mathbf{\Pi}_1$)	15.4200	57.3900	19.8033	10.0800	13.4533	11.9433
Non-players ($\mathbf{\Pi}_2$)	15.5800	57.7700	19.8100	10.9467	13.6967	11.8033

College players ($\mathbf{\Pi}_1$)	WDIM	CIRCUM	FBEYE	EYEHD	EARHD	JAW
WDIM	0.4065	0.6188	0.1954	−0.2320	0.1127	0.2553
CIRCUM	0.6188	2.9313	0.9431	0.1970	0.0930	0.3132
FBEYE	0.1954	0.9431	0.5521	−0.0634	−0.0005	0.1281
EYEHD	−0.2320	0.1970	−0.0634	1.1520	0.0870	−0.1570
EARHD	0.1127	0.0930	−0.0005	0.0870	0.5702	−0.0079
JAW	0.2553	0.3132	0.1281	−0.1570	−0.0079	0.3770

Non-players ($\mathbf{\Pi}_2$)	WDIM	CIRCUM	FBEYE	EYEHD	EARHD	JAW
WDIM	0.3334	0.5746	0.1068	0.2506	0.0851	0.1821
CIRCUM	0.5746	2.3911	0.7000	0.9846	0.0664	0.4867
FBEYE	0.1068	0.7000	0.3802	0.0833	−0.0272	0.1162
EYEHD	0.2506	0.9846	0.0833	1.4577	0.3171	0.1091
EARHD	0.0851	0.0664	−0.0272	0.3171	0.3921	−0.0469
JAW	0.1821	0.4867	0.1162	0.1091	−0.0469	0.2714

Using the stochastic representation in Theorem 3.16 for the discriminant function $\hat{Q}_1$ in (1.8), with $c_1 = c_2 = \frac{1}{2}$, and with all unknown parameters estimated through (5.1) and (5.2), we performed a Monte Carlo simulation to estimate the probabilities of misclassification. The estimates obtained for $P(1 \mid 2)$, $P(2 \mid 1)$ and TPMC were 0.3159, 0.2911 and 0.3036, respectively. As expected, the estimates of $P(1 \mid 2)$ and $P(2 \mid 1)$ are very close to each other.

REFERENCES

1. T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," 2nd ed., Wiley, New York.
2. M. S. Bartlett and N. W. Please, Discrimination in the case of zero mean differences, *Biometrika*, **50** (1963), 17–21.
3. A. H. Bowker, A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistics, *in* "Studies in Item Analysis and Prediction" (H. Solomon, Ed.), pp. 285–292, Stanford Univ. Press, Stanford, CA, 1961.
4. M. Desu and S. Geisser, Methods and applications of equal-mean discrimination, *in* "Discriminant Analysis and Applications" (T. Cacoullos, Ed.), pp. 139–159, Academic Press, New York, 1970.

5. K. R. Gabriel, The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58** (1971), 453–467.
6. C. S. Herz, Bessel functions of matrix argument, *Ann. of Math.* **61** (1955), 474–523.
7. N. L. Johnson and S. Kotz, "Continuous Univariate Distributions-2," Houghton–Mifflin, Boston, 1970.
8. R. A. Johnson and D. W. Wichern, "Applied Multivariate Statistical Analysis," 3rd ed., Prentice–Hall, NJ, 1992.
9. R. Khattree and D. N. Naik, "Applied Multivariate Statistics with SAS® Software," SAS Institute, Cary, NC, 1996.
10. V. R. Marco, D. M. Young, and D. W. Turner, Asymptotic expansions and estimation of the expected error rate for equal-mean discrimination with uniform covariance structure, *Biometrical J.* **29** (1987), 103–113.
11. K. V. Mardia, Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhyā*, *Ser. B*, **36** (1974), 115–128.
12. H. R. McFarland III, "The Exact Distributions of 'Plug-in' Discriminant Functions in Multivariate Analysis," Ph.D. dissertation, University of Virginia, 1998.
13. G. J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition," Wiley, New York, 1992.
14. R. J. Muirhead, Latent roots and matrix variates: a review of some asymptotics, *Ann. Statist.* **6** (1978), 5–33.
15. R. J. Muirhead, "Aspects of Multivariate Statistical Theory," Wiley, New York, 1982.
16. M. Okamoto, Discrimination for variance matrices, *Osaka Math. J.* **13** (1961), 1–39.
17. A. C. Rencher, "Methods of Multivariate Analysis," Wiley, New York, 1995.
18. M. Siotani, Large sample approximations and asymptotic expansions of classification statistics, *in* "Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics" (P. R. Krishnaiah and L. N. Kanal, Eds.), Vol. 2, pp. 61–100, North-Holland, New York, 1984.
19. P. Stocks, A biometric investigation of twins, Part II, *Ann. Eugen.* **5** (1933), 1–55.
20. G. N. Watson, "A Treatise on the Theory of Bessel Functions," Cambridge Univ. Press, New York, 1922.
21. D. M. Young, J. W. Seaman, and L. W. Jennings, Bounds on the error rate for equal-mean discrimination with known population parameters, *Statistica* **47** (1987), 85–96.